

THE IMPACT OF MULTIPLE IMPUTATIONS ON
THE ESTIMATION OF COEFFICIENT ALPHA

By

HON KEUNG YUEN

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2000

ACKNOWLEDGMENTS

I am indebted to a few special individuals who made this dissertation possible. First, I would like to thank my wife, Kit, for her patience and understanding throughout this process. Also, I would like to thank my committee members Dr. David Miller, and Dr. Anne Seraphine, Dr. Kay Walker, and Dr. Arthur Newman for their time and support.

TABLE OF CONTENTS

	<u>page</u>
ACKNOWLEDGEMENTS	ii
ABSTRACT	iv
 CHAPTERS	
1. INTRODUCTION	1
Statement of the Problem	4
Rationale for the Study	5
Purpose and Significance of the Study	6
2. REVIEW OF LITERATURE	7
Common Missing Data Treatments	7
Multiple Imputation	14
Missing Data Mechanisms	33
3. METHODOLOGY	45
Simulation Procedure	46
Design of Study	50
Multiple Imputation Procedure	61
Evaluating the Performance of Multiple Imputation	64
4. RESULTS	66
5. DISCUSSION	79
Limitations	81
Suggestions to Future Research	81
REFERENCES	84
BIOGRAPHICAL SKETCH	90

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

THE IMPACT OF MULTIPLE IMPUTATIONS ON
THE ESTIMATION OF COEFFICIENT ALPHA

By

Hon Keung Yuen

August, 2000

Chairpersons: M. David Miller and Anne Seraphine
Major Department: Educational Psychology

The purpose of this dissertation is to investigate the accuracy of coefficient alpha on tests when nonrandom missing data are replaced using multiple imputation under a single-facet crossed model. The performance of multiple imputation was evaluated under the conditions of three sample sizes ($N = 50, 100, \text{ or } 500$), ten conditions of distribution and percent of missingness, and two omitting patterns (omitting item responses in the body and omitting responses at the end of the test). The ten missing conditions were formed from examinees of the three ability levels (high, medium and low) with a differential number of missing items. The nonrandom nature of missingness leads to examinees with low ability missing more difficult items or more items at the end of the test than those with high ability. A twenty-item test was used in this study. Results of the one thousand iterations indicated that the magnitude of the bias obtained in the omitting pattern where missing responses are at the end of the test was less than 0.03. In contrast, the magnitude of the bias obtained

in the omitting pattern where missing responses are in the body of the test was less than 0.07. In general, the bias increased as the amount of missingness increased or as the sample size decreased. However, this pattern is not uniform across all the missing conditions investigated. Overall, this simulation study confirmed that multiple imputation is a reasonably good procedure to replace the missing data on tests in which missing responses are either in the body of the test or at the end of the test.

CHAPTER 1 INTRODUCTION

Accurate measurement of examinees' ability in standardized achievement assessments requires the test scores to be reliably measured. Internal consistency is one type of reliability that indicates how strongly the test items within the same construct are correlated. Internal consistency of a test appeals to educators because it requires only a single administration of one form of a test. Coefficient alpha (Cronbach, 1971) is a commonly used index to estimate the internal consistency of a test. The index is not a direct estimate of the theoretical reliability coefficient but is an estimate of the lower bound of the internal consistency (Crocker & Algina, 1986). According to Peterson (1994), the formula for computing coefficient alpha (α) can be expressed as

$$\alpha = \frac{s}{s-1} \left[1 - \frac{\sum_{i=1}^s \sigma_i^2}{\sigma_x^2} \right] = \frac{s}{s-1} \left[1 - \frac{\sum_{i=1}^s \sigma_i^2}{\sum_{i=1}^s \sigma_i^2 + 2 \sum_{i < s} \sigma_i \sigma_s r_{is}} \right] \quad (1-1)$$

where

s is the number of items in the test,

σ_x^2 is the variance of the test scores,

σ_i^2 is the variance of a single item i ,

$\sigma_i \sigma_s r_{is}$ is the covariance between item i and item s , and

r_{is} is the correlation between item i and item s .

$$\text{Or } \alpha = \frac{s\bar{r}}{1 + \bar{r}(s-1)} \quad (1-2)$$

where \bar{r} is the average inter-item correlation.

As in the estimation of the Pearson product moment correlation coefficient, computation of coefficient alpha requires a rectangular person-by-item data matrix with no missing data (i.e., a balanced design data set). However, it is well known that missing data is common in large-scale standardized educational achievement tests such as the National Assessment of Educational Progress (NAEP) (Koretz, Lewis, Skewes-Cox, & Burstein, 1993; Longford, 1994) and the Test of English as a Foreign Language (TOEFL) (Yamamoto, 1995). Yamamoto (1995) indicated that about 20% of examinees have difficulty completing the last 20% of the items in the TOEFL. Two main classes of nonrandom omitting pattern in a test have been identified: omitting item responses in the body of the test and omitting item responses at the end of the test (i.e., not-reached) (Longford, 1994). A number of conceivable circumstances can contribute to these occurrences. Angoff and Schrader (1984) found that response omissions in the body of the test is common in tests with instructions indicating that there is a penalty for incorrect responses, but not for response omissions. In this situation, examinees are more likely to omit difficult items when they are not sure of the answer (Koretz et al., 1993). For omitting responses at the end of the test, time constraints is a major factor. However, item difficulty has been reported to contribute to this type of omitting pattern (Koretz et al., 1993). Cluxton and Mandeville (1979) found less capable students tend to omit more items at the end of the test.

Because of the balanced design requirement in the data set to compute coefficient alpha, missing data present a challenge when standard methods of data analysis are used. In the last few decades, a number of missing data treatments (MDTs) have been proposed (see review in Little & Rubin, 1987). A promising MDT is multiple imputation (MI), which was originally proposed by Rubin (1987). MI is a model-based estimation technique for analyzing data with missing scores (Rubin, 1987). Using information from the observed part of the data set, MI generates k sets of equally plausible values from the simulated distribution of the missing data to replace the missing scores, where k is greater than one. The missing scores are imputed k times (Rubin, 1987). As a result, MI creates k versions of complete data sets with imputed values. Each complete data set can be analyzed separately by means of standard complete-case analysis methods. The final adjusted point estimate is obtained by averaging over the k intermediate parameter estimates. MI has been shown to yield satisfactory parameter estimates with relatively little bias (Graham & Schafer, 1999). However, MI has not been used widely in educational settings except for matrix sampling and scaling procedures in the NAEP (Mislevy, Johnson, & Muraki, 1992; Neal & Nianci, 1997).

Several recent studies compared different MDTs in estimating reliability coefficients on measures with missing data (Downey & King, 1998; Harrison, 1998; Marcoulides, 1990). Downey and King (1998) compared the accuracy of coefficient alpha estimation using item mean and person mean substitution to replace missing data in Likert scales. Results indicated that item-mean substitution reduces the reliability estimate whereas person-mean substitution increases the reliability estimate of the scale as the number of missing items and the number of respondents with missing items

increases beyond 20% (Downey & King, 1998). Marcoulides (1990) compared the consistency and efficiency of two MDTs (restricted maximum likelihood and analysis of variance) in estimating variance components on measures with missing data. He found that restricted maximum likelihood (REML) produces a more efficient and less biased variance estimate when 20% of the data are randomly deleted (Marcoulides, 1990). Along the same line of research, Harrison (1998) evaluated six MDTs (listwise deletion, zero imputation, substituting least square ANOVA, substituting probabilities of correct answers from logistic regression estimates, Hoyt's ANOVA formula, and REML) in estimating coefficient alpha on tests with dichotomously-scored items under the conditions of five random and nonrandom missing data patterns crossed with two sample sizes (50 and 100). Results showed that REML provides reasonable accuracy and precision for the estimation of coefficient alpha in all five missing data patterns (Harrison, 1998).

Statement of the Problem

Results of Harrison's (1998) study indicated that the average bias of the coefficient alpha when using each of the six MDTs is negligible (less than 0.05) except in two nonrandom missing-data situations where a listwise deletion procedure is used. One reason for the small discrepancy in the bias among the six MDTs is that the maximum amount of missing data in Harrison's study is less than 11%. Roth (1994) cited several simulated and empirical MDT studies indicating that there is little difference in parameter estimates when the amount of missing data is less than 5-10% regardless of the missing data patterns (random or nonrandom). Roth (1994) suggested that the choice of MDTs

becomes more important when the amount of missing data in a data set is beyond 15-20%. Therefore, we still do not know how some of the MDTs behave in situations with a moderate amount of missingness.

Harrison (1998) found that the mean coefficient alpha produced by REML is more positively biased (i.e., overestimated) than that computed by Hoyt's ANOVA in situations where low-ability examinees have more omitted items or where they tend to omit the most difficult items. There are two possible reasons for REML not behaving well in Harrison's study. One is that REML has been shown to produce estimates that are significantly biased in situations where sample size is small ($N = 20$ or 50) because REML is based on large-sample theory (Gross, 1997). The other is that REML produces biased estimates when data are missing nonrandomly (Jamshidian & Bentler, 1999).

Rationale for the Study

The present study attempts to address some of the limitations of the REML estimation procedure (Harrison, 1998) by implementing MI which has been shown to perform well in small sample sizes (Graham & Schafer, 1999) and in nonrandom missing-data situations (Graham, Hofer, Donaldson, MacKinnon, & Schafer, 1997). Although MI is commonly applied to missing continuous data, it may also be applied to dichotomous missing data (Graham et al., 1997). Because Harrison's study examined the effectiveness of different MDTs under slight levels of missingness, it is of interest to examine the MI under more extreme levels of missingness. The level of missingness is therefore set as high as 30%. At this range of missingness, it would become more obvious how well MI performs.

It is well known that data missing completely at random (MCAR) seldom occurs in educational settings (Kromrey & Hines, 1994). The present study, therefore, focuses on nonrandom missing data.

Purpose and Significance of the Study

The purpose of this study was to investigate, via data simulation, the accuracy of the coefficient alpha on tests with missing data replaced using MI. The performance of MI was evaluated under the conditions of three sample sizes ($N = 50, 100, \text{ or } 500$), ten conditions of distribution and percent of missingness, and two omitting patterns (omitting item responses in the body and omitting responses at the end of the test). The results of this study provided an indication of how well MI performed in the above-stated missing data conditions under a single-facet crossed model.

CHAPTER 2

REVIEW OF RELATED LITERATURE

The first section of this chapter provides an overview of some commonly used missing data treatments (MDTs), which include listwise deletion, variable mean substitution, regression imputation, and stochastic regression imputation. Limitations of these MDTs are highlighted. The second section is devoted to the development and theoretical framework of multiple imputation (MI), the relationship of MI and Bayes' theorem, assumptions and characteristics of MI, the description of the imputation methods, and procedures to perform MI. The last section discusses three major types of missing data mechanisms as proposed by Little and Rubin (1987), and implications of each missing data mechanism to the application of MI.

Common Missing Data Treatments

Listwise Deletion

In order to transform the missing data matrix into a rectangular one, a common practice is to exclude those examinees who do not respond to all items. This is called the listwise deletion procedure or the complete-case analysis. The complete data with reduced sample size are then used to estimate population parameters such as the reliability coefficient. Listwise deletion is the default option for analysis in many popular statistical software packages such as the Statistical Analysis System (SAS) and the Statistical Packages for the Social Sciences (SPSS-X). Even though listwise deletion is

the simplest approach to handle missing data, it is by no means the desirable one. Since the analysis is based on only those examinees who respond to all items, a substantial amount of useful data is lost. In a Monte Carlo investigation, Kim and Curry (1977) found that even with 2% random nonresponses on each of the 10 variables, listwise deletion results in retaining only 81.7% of the cases. There is an accompanied loss of efficiency or statistical power in the estimation of the population parameters especially when the amount of missing data is high. Raaijmakers (1999) demonstrated that listwise deletion results in a loss of statistical power ranging from 35% to 98% as the amount of missing data increases from 10% to 30% in various Likert-type data.

Listwise deletion is based on the assumption that the data are missing completely at random even though there is little evidence to support this assumption in educational research (Kromrey & Hines, 1994). When data are not missing completely at random, estimates are biased. Empirical data support that the average bias increases as the amount of missing data increases (Harrison, 1998). In Harrison's (1998) study, when item responses are not missing at random, listwise deletion leads to range restriction. The resulting mean coefficient alpha is then substantially underestimated (Harrison, 1998).

Single Imputation Procedures

Besides using deletion procedures to handle missing data, single imputation procedures have also been used widely in educational research (see review in Raymond, 1987). Imputation involves filling in each missing response with a plausible value and then analyzing the resulting data set with the imputed values. Also, plausible values are estimated from observed scores in the study. Two major advantages of imputation procedures are as follows:

1. They retain the information from incomplete cases without discarding any scores.
2. The resulting data set with the imputed values can be analyzed by means of standard complete-case analysis methods.

The three single imputation procedures that are commonly used in educational research are variable mean substitution, regression imputation, and stochastic regression imputation (Raymond, 1987; Roth, 1994).

Variable Mean Substitution

To implement variable mean substitution or unconditional mean imputation, each missing score of a particular item is replaced with its respective mean value of all nonmissing cases. Even though it seems that the mean is a good estimate and the procedure is relatively easy to implement, variable mean substitution has several serious disadvantages. The observed variance of an item with imputed mean value is systematically underestimated (i.e., negatively biased) because imputing a mean value in an item is equivalent to adding zero to the sum of the squared deviations, which is the numerator of the formula for calculating variance. At the same time, there is an increase in the denominator of the variance formula, $(N - 1)$, as the procedure attempts to restore the original sample size (Landerman, Land, & Pieper, 1997; Raymond, 1987).

Attenuation of the magnitude of the covariance or correlation of scores with filled-in mean with scores in other items can be explained in a similar fashion. Conceptually, the imputed values are a constant and they are unrelated to scores in other items; therefore, inter-item correlations are attenuated. Downey and King (1998) showed that the severity of attenuation on correlation increases as the amount of imputed values increases. As indicated in equation (1-2), a reduction in the average inter-item correlation

results in a decrease in the coefficient alpha (Downey & King, 1998). Figure 2-1 graphically shows the coefficient alpha spuriously decreases when there is a reduction in the average inter-item correlation at the lower end (i.e., $\bar{r} < 0.2$).

Another disadvantage related to the attenuation of variability of the item is that the standard error of estimate is much too small resulting in biased inferences (Little & Rubin, 1989). Finally, variable mean substitution does not use information from other items to improve the accuracy of imputation (Landerman et al., 1997).

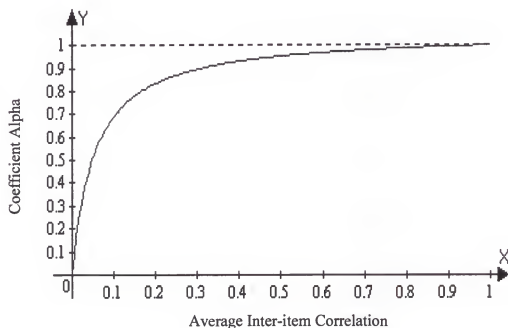


Figure 2-1. Relationship between the coefficient alpha and the average inter-item correlation when s equals to 20.

Regression Imputation

Regression imputation or conditional mean substitution is used to fill in missing scores of an item with values predicted from a regression model by utilizing information from one or more highly related observed variables or predictors. When the response variable with missing scores is dichotomous in nature, a logistic regression model is used instead. The logistic regression produces a predicted probability of a response being missing.

Suppose Y is an $N \times 1$ vector of the responses for examinees, $Y = (y_1, \dots, y_n)$ and is composed of a set of the observed scores, $Y_o = (y_1, \dots, y_o)$ and a set of the missing scores, $Y_m = (y_{o+1}, \dots, y_n)$. Y can be partitioned into $Y = (Y_o, Y_m)$. Let n_o be the number of respondents for Y_o and n_m be the number of nonrespondents for Y_m . X denotes an $N \times G$ design matrix of relevant explanatory variables or predictors that are highly correlated with Y . These variables are fully observed (i.e., with no missing data). X_o represents the predictors for n_o individuals with observed scores Y_o and X_m represents the predictors for n_m individuals with missing scores Y_m . Y_m scores are to be estimated. A linear regression equation can be expressed as

$$Y = a + bX + e \quad (2-1)$$

where

a is a column vector of the intercepts,

b is a column vector of estimated regression coefficients, and

e is a column vector of estimated residuals and is set to zero.

The procedure to impute the predicted values based on a deterministic regression model is as follows. First, use X_o to estimate the coefficients of the linear regression

equation by regressing Y_o on \mathbf{X}_o . After estimating the regression coefficients from the observed scores, a predicted score for Y_m can be obtained from the prediction equation:

$$\hat{Y}_m = a + \hat{b}\mathbf{X}_m \text{ (Little \& Rubin, 1989).}$$

Landerman and associates (1997) explained that the distribution of Y based on the regression imputation is less distorted than Y based on the mean substitution because the imputed values are now distributed across the predicted values \hat{Y}_m instead of concentrated at the mean. The variance of Y based on the regression imputation is less attenuated than that based on the mean substitution because the numerator for calculating variance is the squared deviations of the \hat{Y}_m from the grand mean, which is not likely to be zero.

In regression imputation, the imputed values of Y_m fall exactly on the predicted regression line or plane as the estimated residual e in equation (2-1) is set to zero (Landerman et al., 1997; Little & Rubin, 1989). Therefore, there is no variation in the distribution of Y_m given \mathbf{X}_m (Little & Rubin, 1989). Because of the exact linear (or plane) relationship between Y_m and \mathbf{X}_m , its correlation is spuriously inflated (Graham & Schafer, 1999). Harrison (1998) demonstrated that replacing missing data with either the least squares estimates or the predicted probabilities overestimates the coefficient alpha. This phenomenon can be explained by the positive relationship between the coefficient alpha and the average inter-item correlation in equation (1-2). In addition, the severity of bias (overestimation of the coefficient alpha) increases as the amount of missing data increases (Harrison, 1998). Little (1992) concluded that mean substitution or regression imputation can yield unbiased estimates of aggregate means but leads to distorted variance and covariance estimates.

Regression imputation conveys a false sense of accuracy that all missing scores can be predicted from X_m without errors. By treating the imputed values as the known observed scores, regression imputation fails to account properly for the variability or uncertainty about not knowing the missing scores (i.e., which value to impute) (Rubin & Schenker, 1991). Failure of the regression model to incorporate residual variability in the imputation variance leads to standard errors of estimates bias toward zero (i.e., too small) (Little & Rubin, 1989). For example, Brownstone and Valletta (1996) found that the least squares standard error estimates are 30% less than their true values.

Stochastic Regression Imputation

In order to restore the prediction errors in the imputed values (i.e., the variability around the regression line), a random residual / error is added to each predicted value. The random residual can be drawn randomly with replacement either from a standard normal distribution with a mean equal to zero and a standard deviation equal to the standard error of estimate for Y_o (Beaton, 1997), or from the distribution of residuals of the regression estimate for Y_o (Graham et al., 1997). The purpose of drawing with replacement is to ensure that each drawn value has equal probability. In stochastic regression imputation, each missing response is replaced by its conditional mean plus a random residual from Y_o (Little & Rubin, 1989).

However, stochastic regression imputation restores only one part of the variability: the errors of prediction. There is another part of variability, the sampling variability in which the values of the estimated regression coefficients are uncertain. Graham and Schafer (1999) explained that the regression line estimated from Y_o is not the regression for the population, but is only an estimate from one sample. Stochastic

regression imputation cannot reflect properly the sampling variability because it lacks the variation of the imputed values among several sets of imputations (Little & Rubin, 1989). To incorporate the sampling variability in the estimation of the regression parameters, multiple imputation (MI) is required (Rubin, 1987).

Multiple Imputation

Introduction

Multiple imputation was originally proposed by Rubin (1987). It is a model-based estimation technique for analyzing data with missing scores (Rubin, 1987). Using information from the observed part of the data set, MI generates k sets of equally plausible values from the simulated distribution of the missing data to replace the missing scores, where k is greater than one. The missing scores are imputed k times and multiple imputations within one model are called repetitions (Rubin, 1987). As a result, MI creates k versions of complete data sets with imputed values. Each complete data set can be analyzed separately by means of standard complete-case analysis methods. The estimate and its associated variance from each separate analysis can be combined to form an unbiased final parameter estimate under the correctly specified model (Little & Rubin, 1989). The final variance incorporates the variability within the imputation (i.e., the prediction error) and the variation of the imputed values among k sets of imputations (i.e., the sampling variability) to reflect the true accuracy of the estimation.

Theoretical Framework

Let Q denote a scalar population quantity (such as a coefficient α or a regression coefficient) to be estimated and let $Q = Q(Y_o, Y_m)$ denote a function of the

observed and missing data. Multiple imputation uses information from the observed scores Y_o to replace the missing scores Y_m , and then uses the complete data set with imputed values to estimate the parameter Q . A distribution of the missing data is required to generate the imputed values. The distribution is drawn from the observed scores Y_o . It is necessary to know the structural or model parameter of the observed scores, which is θ , where θ represents a vector of q parameters. For example, $\theta = (\mu, \sigma^2)$, means that θ is a function of the mean μ and the variance σ^2 . Because θ is unknown, it must be estimated, resulting in the random variable, $\hat{\theta}$ (Michiels & Molenberghs, 1997). Because of the uncertainty of not knowing θ , Rubin (1987) recommended using Bayesian methodology to account for this uncertainty in MI. Through a Bayesian procedure, a distribution function of $\hat{\theta}$ in the form of the posterior probability distribution of θ can be obtained from the data (Michiels & Molenberghs, 1997).

The derivation of MI is as follows (Rubin, 1996): Inferences for Q are based on the actual posterior probability distribution of Q , $f(Q | Y_o)$, which can be expressed as

$$f(Q | Y_o) = \int f(Q | Y_o, Y_m) f(Y_m | Y_o) dY_m \quad (2-2)$$

where

f denotes the probability distribution function,

$f(Q | Y_o, Y_m)$ is the complete data posterior probability distribution of Q and is expressed as the conditional distribution of Q given both the observed and missing data, and

$f(Y_m | Y_o)$ is the predictive probability distribution of missing scores Y_m given the observed scores Y_o .

Based on equation (2-2), the actual posterior probability distribution of Q at a particular value Q_i can be obtained by drawing an infinite number of repeated independent values for Y_m from $f(Y_m | Y_o)$, calculating $f(Q_i | Y_o, Y_m)$ separately for each draw, and then averaging the values over the repeated imputations (Little & Schenker, 1995; Rubin, 1996).

The predictive distribution of the missing scores can be parameterized using a structural parameter θ (Little & Schenker, 1995), and is expressed as

$$f(Y_m | Y_o) = \int f(Y_m | Y_o, \theta) f(\theta | Y_o) d\theta \quad (2-3)$$

where

$f(\theta | Y_o)$ is the conditional distribution of θ given the observed scores Y_o , and

$f(Y_m | Y_o, \theta)$ is the conditional distribution of Y_m given the observed scores Y_o and the parameter θ .

From the Bayesian perspective, drawing k values for the missing scores Y_m in MI involves two steps (Schafer, 1999):

Step 1. Simulate an independent random draw of the unknown parameter θ^* from the observed-data posterior distribution $f(\theta | Y_o)$.

$$\theta^* \sim f(\theta | Y_o)$$

where θ^* is the posterior distribution of θ or a distribution for the missing scores estimated.

Step 2. Randomly draw missing values Y_m^* from the conditional predictive distribution of Y_m given parameter θ^* .

$$Y_m^* \sim f(Y_m | Y_o, \theta^*)$$

These two steps are repeated k times to yield k sets of imputed values for the missing scores Y_m .

In principle, MI involves k repetitions of independent draws from the posterior predictive distribution of Y_m by specifying a prior distribution of the unknown structural parameter θ (Little & Schenker, 1995). This forms the k imputations for the missing scores Y_m .

Bayes' Theorem

The aim of MI is to estimate the unknown structural parameter θ and to generate the imputed values. Conditional upon a sample of observed scores Y_o , Bayes' theorem makes inferences about the unknown parameter θ . Bayes' theorem represents uncertainty of not knowing θ by a prior probability distribution (Pollard, 1986).

The conditional distribution of θ given the observed scores Y_o or the posterior distribution of θ is derived from a Bayesian procedure (Pollard, 1986) and is defined as

$$f(\theta|Y_o) = f(Y_o|\theta)f(\theta) / f(Y_o) \quad (2-4)$$

where

$f(Y_o|\theta)$ is the conditional probability, or likelihood, of the observed scores Y_o ,

$f(\theta)$ is the prior distribution of unknown structural parameter θ and represents uncertainty about the value of the parameter θ before any data are seen, and

$f(Y_o)$ is the marginal probability of observed scores Y_o for an examinee of parameter θ randomly sampled from a population with the given distribution.

Since $f(Y_o)$ is a constant that serves to make $f(\theta|Y_o)$ integrate to one, the equation (2-4) becomes

$$f(\theta | Y_o) \propto f(Y_o | \theta) f(\theta) \quad (2-5)$$

where \propto indicates a relationship of proportionality.

Given Y_o , $f(Y_o | \theta)$ becomes a likelihood function for θ given Y_o , $L(\theta | Y_o)$. The posterior distribution of θ is then written as

$$f(\theta | Y_o) \propto L(\theta | Y_o) f(\theta) \quad (2-6)$$

where $L(\theta | Y_o)$ expresses the information about the parameter θ provided by the observed scores Y_o , and $f(Y_o | \theta)$ serves to convert the prior distribution $f(\theta)$ into the posterior distribution $f(\theta | Y_o)$ (Pollard, 1986).

It is through the likelihood (i.e., $L(\theta | Y_o)$) of $f(Y_o | \theta)$ that the observed scores Y_o modify the prior distribution $f(\theta)$ to determine the posterior distribution $f(\theta | Y_o)$ (Pollard, 1986). In essence, Bayesian methodology specifies a distribution of what is expected to occur based on prior information and combines it with new information (i.e., the observed scores Y_o) to form inferences about θ .

Assumptions Underlying Multiple Imputation

The missing data mechanism is assumed to be ignorable or missing at random (Graham & Schafer, 1999). Ignorable means it is not necessary to specify a nonresponse model or to estimate its parameters in order to obtain valid likelihood-based inferences. Missing at random means that the missing data are a random sample from the complete data after conditioning on the measured variables X in the imputation model (Schafer, 1997). However, Graham and associates (1997) have demonstrated that MI produces satisfactory parameter estimates even when the ignorability assumption is suspect.

In addition, the variables in the data set are assumed to have multivariate normal distribution. Simulation studies (Graham, Hofer, & McKinnon, 1996; Wang, Anderson, & Prentice, 1999) supported that the MI estimator is robust even when the data model departs from being multivariate normally distributed.

Proper Imputation Method

An imputation method is regarded as proper when it incorporates appropriate variability (i.e., uncertainty about the missing scores and the sampling variability) in creating multiply imputed values under a correctly specified model (Rubin, 1987, 1996). Rubin (1987) has shown that one way to achieve proper imputation is for the imputation procedure to follow the Bayes' theorem of infinite independent draws of Y_m from its posterior predictive distribution as specified in equations (2-2 & 2-3). By incorporating variability to adjust the standard error of estimates of the parameter, proper imputation method leads to valid inferences (Rubin & Schenker, 1991).

The conditions under which an imputation method is proper include the following:

1. Imputed values are independent repeated draws from a Bayesian posterior predictive distribution of the missing scores Y_m given the observed scores, $f(Y_m | Y_o)$ (Rubin, 1996).
2. Infinite k repeated imputations since parameter estimates derived from infinite draws for Y_m are fully efficient (Little & Rubin, 1989).
3. The underlying model specification for the complete data is correct.
4. The underlying model specification for the missing data mechanism (i.e., assumptions about the nonresponse) is correct.

5. Large-sample size ($N > 100$) (Rubin & Schenker, 1986).
6. All causes of missingness are included in the imputation model (Graham et al., 1997).

Imputation Methods

Rubin and Schenker (1986) proposed two types of imputation methods—implicit and explicit. Implicit or nonparametric methods are applicable for discrete data and involve drawing values only from Y_o and then assigning them to Y_m . In contrast, explicit or parametric methods are applicable for continuous data and involve a statistical model to form the posterior predictive distribution of Y_m , from which imputing values are drawn. Unlike implicit methods, the drawing values based on explicit methods are not in Y_o (Rubin & Schenker, 1986).

Implicit Methods

Simple hot deck procedure

The simple hot deck procedure involves random draws with replacement of n_m imputed values for nonrespondents from observed scores in matching respondents. However, like stochastic regression imputation, simple hot deck procedure ignores the sampling variability as the population distribution of $(Y_m | Y_o)$ is not known. The imputed values are estimated from the respondent scores Y_o in one sample only (Little & Schenker, 1995).

Approximate Bayesian bootstrap

In order to incorporate sampling variability in the estimated parameters, approximate Bayesian bootstrap (ABB) is used (Rubin, 1987). The ABB creates k repeated imputations from the posterior predictive distribution of the missing data as follows:

1. Draw n_o values at random with replacement from the n_o possible values to create a bootstrap sample distribution such as a scaled multinomial distribution.
2. Then independently draw n_m missing values with replacement from the bootstrap sample distribution (Rubin & Schenker, 1986).

This process is repeated k times to yield k sets of imputed values, and each set of imputations comes from a different bootstrap sample of Y_o .

Explicit Methods

Explicit methods define the model for the distribution of the response variable Y (e.g., normal linear regression model or logistic regression model) and a set of predictors X that enters the model to create imputations (Little, 1992; Rubin & Schenker, 1991).

Fully normal imputation

Once again, suppose Y is an $N \times 1$ vector of the response for examinees, $Y = (y_1, \dots, y_n)$ and is composed of both a set of the observed scores, $Y_o = (y_1, \dots, y_o)$, and a set of the missing scores, $Y_m = (y_{o+1}, \dots, y_n)$. Let n_o be the number of respondents for Y_o and n_m be the number of nonrespondents for Y_m . The scores of Y_m are to be estimated. Rubin and Schenker (1986) described how to create multiple imputations under the independent normal model, $y_i \sim N(\mu, \sigma^2)$, for $i = 1, \dots, o$, where $\theta = (\mu, \sigma^2)$ is unknown, and θ is a function of the mean μ and the variance σ^2 . When the prior distribution of $\theta, f(\theta)$, is proportional to $1/\sigma^2$, the conditional posterior distribution of μ given $\sigma^2, f(\mu | \sigma^2, Y_o)$ is $N(\bar{y}_o, \sigma^2 / n_o)$,

where \bar{y}_o is the sample mean of Y_o , and is equal to $\frac{1}{n_o} \sum_{i=1}^o y_i$;

and the observed-data posterior distribution of $\sigma^2, f(\sigma^2 | Y_o)$ is

$$(n_o - g)\hat{\sigma}^2 / \chi^2_{n_o - g},$$

where $\hat{\sigma}^2$ is the estimated variance of Y_o , and is equal to $\frac{1}{(n_o - g)} \sum_{i=1}^o (y_i - \bar{y}_o)^2$, and

$\chi^2_{n_o - g}$ denotes a chi-square random variable with $n_o - g$ degrees of freedom.

To create an imputation $Y_m^* = (Y_{o+1}^*, \dots, Y_n^*)$, for $^* = 1, \dots, k$, the following three steps are required.

Step 1. Generate the unknown parameters $\theta^* = (\mu, \sigma^{*2})$ from the observed-data posterior distribution $f(\theta | Y_o)$ by first randomly drawing the variance σ^{*2} from $(n_o - g)\hat{\sigma}^2 / \chi^2_{n_o - g}$, and then randomly drawing the mean μ^* from $N(\bar{y}_o, \sigma^{*2} / n_o)$.

Step 2. Independently draw n_m missing values of Y_m from $y_i^* \sim N(\mu^*, \sigma^{*2})$, for $i = o + 1, \dots, n$ (Rubin & Schenker, 1986, Schafer, 1999).

Step 3. Repeat the procedure k times (i.e., set $^* = k$) to yield k sets of proper imputations.

Markov Chain Monte Carlo

In addition to using the resampling procedure such as bootstrapping, which is a noniterative method for creating posterior predictive probability distribution from Y_o , Markov Chain Monte Carlo (MCMC) is a collection of iterative simulation-based methods for generating the posterior distribution of the unknown parameters θ and they do not require large samples for efficacy (Little & Schenker, 1995). Data augmentation algorithm (Tanner & Wong, 1987) and Gibbs sampling (Gelfand & Smith, 1990) are two of the most common MCMC methods used in MI (Little & Schenker, 1995). They generate simulation-based estimates of the Bayesian posterior predictive distribution of

the missing data, $f(Y_m | Y_o)$ and from it, perform k independent random draws of Y_m (Schafer 1997).

Normal Linear Regression Model

Y is modeled by a linear regression model, $Y \sim N(\mathbf{X}\beta, \sigma^2)$ with a multivariate normal distribution, where $\mathbf{X}\beta$ covariate is a function of the parameters, \mathbf{X} contains g variables, β is the parameter vector of regression coefficients to be estimated, and σ^2 is the regression variance. The algorithm for creating k multiply imputed values involves the following steps (Rubin, 1987):

Step 1. Regress Y_o on \mathbf{X}_o to give the ordinary least squares estimates: estimated regression coefficient vector $\hat{\beta}$ and estimated regression variance $\hat{\sigma}^2$.

$$\hat{\beta} = \mathbf{V}\mathbf{X}'_o Y_o \quad (2-7)$$

where $\mathbf{V} = (\mathbf{X}'_o \mathbf{X}_o)^{-1}$.

The vector of predicted responses:

$$\hat{Y}_o = \mathbf{X}_o \hat{\beta} \quad (2-8)$$

The maximum likelihood estimator of σ^2 :

$$\hat{\sigma}^2 = (Y_o - \hat{Y}_o)^2 / (n_o - g) \quad (2-9)$$

This is the estimation task for the normal linear regression model (Rubin, 1987).

Imputation task for this model is from Steps 2 to 4.

Step 2. Estimate σ^{*2} (square of a random error) to account for the deviations around the regression line.

$$\sigma^{*2} = \hat{\sigma}^2 (n_o - g) / L \quad (2-10)$$

for $\ast = 1, \dots, k$; where L is a randomly drawn variate from a chi-squared distribution with $n_o - g$ degrees of freedom.

Substitute $\hat{\sigma}^2 = (Y_o - \hat{Y}_o)^2 / (n_o - g)$ in equation (2-10), and $\sigma^{\ast 2}$ becomes $(Y_o - \hat{Y}_o)^2 / L$.

Step 3. Estimate the regression coefficients by adding the random error σ^{\ast} to account for the uncertainty about the regression prediction.

$$\beta^{\ast} = \hat{\beta} + \sigma^{\ast} V^{1/2} Z \quad (2-11)$$

where

Z is a g -component vector of standard normal deviate, $Z \sim N(0, I_g)$,

I_g is the identity matrix of order g .

Z is formed by drawing g independent variates from $N(0,1)$,

$\sigma^{\ast} V^{1/2}$ is the mean square error of the regression equation, and

$\sigma^2 V$ is the variance-covariance matrix Σ . The roots of the main diagonal of Σ are the standard errors,

$V^{1/2}$ is the triangular square root of V obtained from the Cholesky decomposition, and

$\sigma^{\ast} V^{1/2} Z$ represents the error term.

Each set of randomly drawn coefficients is then used to estimate the missing values, and different sets of coefficients reflect the variation of the regression lines due to sampling. Steps 2 and 3 indicate random draws from the posterior distribution of β (van Buuren, Boshuizen & Knook, 1999).

Step 4. Predict the missing values of Y_m based on the following equation:

$$Y_m^{\ast} = X\beta^{\ast} + \sigma^{\ast} z \quad (2-12)$$

where z is a random number drawn from normal deviates.

Each set of predicted values for Y_m is based on a different set of regression predictions and random components σ^* .

Step 5. Repeat steps 2 to 4 k times to create k sets of imputed values $Y_m^1, Y_m^2, \dots, Y_m^k$.

Repeated-Imputation Inferences

The final point estimate of the parameter Q approximates the actual posterior mean of Q , $E(Q | Y_o)$. This equals the average of the repeated complete-data posterior means of Q , and can be expressed as (Rubin, 1996):

$$E(Q | Y_o) = E[E(Q | Y_o, Y_m) | Y_o] \quad (2-13)$$

where E refers to the expectation over the repeated imputations, and $E(Q | Y_o, Y_m)$ approximates the complete-data posterior mean.

The final estimated variance of the parameter Q approximates the actual posterior variance of Q , $\text{Var}(Q | Y_o)$. This equals the average of the repeated complete-data posterior variances of Q plus the variance of the repeated complete-data posterior means of Q , and can be expressed as (Rubin, 1996):

$$\text{Var}(Q | Y_o) = E[\text{Var}(Q | Y_o, Y_m) | Y_o] + \text{Var}[E(Q | Y_o, Y_m) | Y_o] \quad (2-14)$$

where Var refers to the variance over the repeated imputations, and $\text{Var}(Q | Y_o, Y_m)$ approximates complete-data posterior variance.

Based on the above standard probability derivations (2-13 & 2-14), the k point estimates and their associated variances obtained from the standard complete-case analysis method can be combined into a final adjusted estimate and its estimated variance using Rubin's formulas (Rubin, 1987).

After generating k imputed data sets using an appropriate imputation model and method, and analyzing each of them separately using a standard complete-case analysis

method, MI yields k intermediate parameter estimates $\hat{Q}_i: (\hat{Q}_1, \dots, \hat{Q}_m)$, and k associated variance estimates $\hat{U}_i: (\hat{U}_1, \dots, \hat{U}_k)$, for $i = 1, \dots, k$. The final adjusted point estimate \bar{Q} is obtained by averaging over the k intermediate parameter estimates, which is

$$\bar{Q} = \frac{1}{k} \sum_{i=1}^k \hat{Q}_i \quad (2-15)$$

Whereas the final estimated total variance is obtained by the sum of the average associate variance within a set of k imputed values and the variance across independent sets of imputed values, which is

$$T = \bar{U} + (1 + k^{-1})B \quad (2-16)$$

where \bar{U} is the average within-imputation variance within a set of k imputed values, and is expressed as

$$\bar{U} = \frac{1}{k} \sum_{i=1}^k \hat{U}_i \quad (2-17)$$

and B is the variance across independent sets of imputed values, and is expressed as

$$B = \frac{1}{k-1} \sum_{i=1}^k (\hat{Q}_i - \bar{Q})^2 \quad (2-18)$$

Bakik, Murhy, and Anthony (1998) indicated that the within-imputation variance is a measure of the uncertainty about not knowing the missing data and the between-imputation variance is a measure of ordinary sampling variation. The inflation factor $(1+k^{-1})$ accounts for the simulation errors in using a finite number of imputations (i.e., $k < \infty$) (Barnard & Meng, 1999). Multiple imputation correctly adjusts the standard error of estimates of the parameter by including within- and between-imputation variances.

When there are no missing data, $\hat{Q}_1, \dots, \hat{Q}_k$ are identical, and the between-imputation variance B becomes zero and T is equal to \bar{U} . When $k = 1$ (i.e., single imputation), B cannot be estimated. T is then equal to \bar{U} , and the variance is systematically underestimated (Heitjan & Rubin, 1990). As k increases, both \bar{Q} and T decrease, hence resulting in greater precision of sample statistics (Little & Schenker, 1995).

The extent of influence of missing data on the estimation of Q is determined by both γ and r . The factor r estimates the proportional increase in variance due to missing data, and can be expressed as

$$r = (1 + k^{-1})B / \bar{U} = \gamma / (1 - \gamma) \quad (2-19)$$

where the ratio of B to \bar{U} is a reflection of how much information in the missing part of the data relative to the observed part (Schafer & Olsen, 1988), and γ is an estimate of the fraction of missing data about Q (Little & Schenker, 1995). Little and Rubin (1989) pointed out that γ is equal to the fraction of data missing only when the missing data mechanism is missing completely at random.

Uncertainty

Since the imputed values are not the true observed scores, MI takes into account the uncertainty about the true values of the missing scores in the parameter estimates by drawing parameter θ^* from the observed-data posterior distribution $f(\theta | Y_o)$ and then imputed values Y_m^* from the conditional predictive distribution of Y_m given that parameter θ^* , $f(Y_m | Y_o, \theta^*)$ (Rubin & Schenker, 1991).

In addition to incorporating the uncertainty about not knowing the missing scores, MI also takes into account the fact that the population distribution of Y_m given Y_o is not known, and is estimated from the observed scores Y_o in one sample (Graham & Schafer, 1999; Little & Schenker, 1995). The variation in estimating the regression line is called sampling variability (Graham & Schafer, 1999). The third source of uncertainty / variability comes from the finite number of imputations derived from using approximations to Bayesian posterior distributions, and is called simulation errors (Rubin & Schenker, 1991). Finally, by comparing parameter estimates across a number of plausible missing-data models (i.e., sensitivity analysis), MI reveals uncertainty about reasons for nonresponse (Beaton, 1997; Little & Rubin, 1989).

Number of Imputations

Under the ignorable response assumption, the final adjusted point estimate \bar{Q} and its estimated variance based on infinite number of imputations are the same as the ones obtained from the maximum likelihood estimation (MLE), which is fully efficient and correct (Little, 1992). The large-sample efficiency of the point estimate \bar{Q} based on k imputations relative to that based on an infinite number of imputations is $\sqrt{1 + (\gamma/k)}$, in standard error units (Rubin, 1996; Schafer & Olsen, 1988). As illustrated, with 30% missing data ($\gamma = 0.3$), an estimate based on $k = 3$ imputations has a standard error of $\sqrt{1 + 0.3/3} = 1.05$. This means the standard error is 5% wider than the one obtained from MLE. Alternatively, the percent efficiency of \bar{Q} is defined as $1/\sqrt{1 + (\gamma/k)}$ (Rubin, 1987). In this example, the percent efficiency is $1/1.05 = .95$ or 95%. This means the efficiency of \bar{Q} is 5% less than the one obtained from MLE. By increasing k to 5 and 10

imputations, it increases the efficiency of \bar{Q} to 97% and 99.5% respectively. As shown in Figure 2-2, unless the fraction of missing data is unusually high (70% or more), the efficiency gained by implementing k beyond 5-10 is minimal. Rubin and Schenker (1986) concluded that only a few number of repetitions ($3 \leq k \leq 10$) are needed to produce point estimates that are close to fully efficient when the amount of missing data is moderate (e.g., 30%).

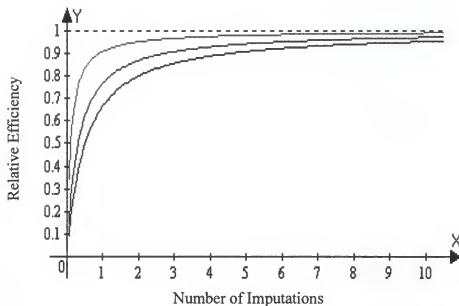


Figure 2-2. Percent efficiency of MI estimation using different number of imputations in three levels of missingness (10%, 30%, and 50%).

Several empirical studies supported that standard error estimates of the parameters were underestimated by 10-20% in single imputation when compared to the ones in MI (Crawford, Tennstedt, & McKinlay, 1995; Heitjan & Rubin, 1990; Landerman et al., 1997). Based on the results of two Monte Carlo simulation studies (Little & Rubin, 1989; Rubin & Schenker, 1986), Table 1 shows the comparison of the actual confidence interval (CI) coverage for \bar{Q} , when $k = 1, 2$, or 3 , with the nominal coverage at 90%, 95%, or 99%. Under the ignorable response assumption, the fraction of missing data in these two large-sample ($N > 100$) simulation studies was 30%. As indicated, when $k = 1$ (single imputation), the discrepancy between the actual and nominal coverage ranges from 5-13%, whereas when $k = 2$, the discrepancy is only 2-3%. When $k = 3$, there is no discrepancy at all, which means that inferences are valid.

Table 2-1. Analytic Large-Sample ($N > 100$) Coverage (in %) of Single ($k = 1$) and Multiple ($k = 2$ or 3) Imputation Procedure with Missing Data Equals 30%

k	Nominal Coverage		
	90%	95%	99%
1	77	85	94
2	87	93	99
3	90	95	99

Note. Adapted from Little and Rubin (1989) and Rubin and Schenker (1986)

Rubin and Schenker (1986) demonstrated that k should increase from 2 to 3 as the nonresponse rate increases from 10% to 60% in order to achieve a satisfactory CI coverage (i.e., close to the nominal value). Rubin and Schenker (1986) also pointed out that improvements in the actual CI coverage diminish as k increases. The differences in standard errors or CI coverage between $k = 5$ and $k \geq 25$ have been shown to be negligible (Heitjan & Rubin, 1990; Wang, Sedransk, & Jinn, 1992).

Based on the literature, the number of imputations depends on:

1. The amount of missing information. As the percent of missing data increases, the amount of uncertainty about the imputed values increases. To accurately incorporate this uncertainty, it requires an increase in the number of imputations. Since imputed values are averaged over k imputations, the imputation variance is reduced as the number of imputations increases (Kalton & Kasprzyk, 1986; Rubin, 1987).
2. The type of missing data mechanisms. Based on several simulation and empirical studies, Glynn, Laird, and Rubin (1993) and Raghunathana and Siscovick (1996) demonstrated that nonignorable nonresponse patterns require a larger number of imputations ($k > 10$) than ignorable nonresponse patterns to achieve a satisfactory CI coverage.

Advantages of Multiple Imputation

As in single imputation, the resulting data set with the imputed values can be analyzed by means of standard complete-case analysis methods. Because MI involves averaging over k intermediate parameter estimates, the final point estimate derived from MI is more efficient than that from single imputation (Rubin, 1996). The final estimated variance also reflects the true variance of the parameter.

Studies affirmed that MI produces accurate standard errors (i.e., efficient) for parameter estimates as it correctly adjusts for nonresponse bias (Heitjan & Little, 1991; Rubin & Schenker, 1986; Xie & Paik, 1997). The estimated actual CI coverage is close to the nominal levels (Little & Rubin, 1989; Rubin & Schenker, 1986; Wang et al., 1992), which means MI yields valid inferences.

MI has been shown to yield satisfactory parameter estimates with relatively little bias even under the following conditions:

1. Sample sizes are small (e.g., 50) (Graham & Schafer, 1999). Little (1992) recommended to use MI for small samples and MLE for large samples.
2. Data are missing in large amounts (e.g., 50%) (Graham & Schafer, 1999).
3. Models are relatively large and complex (e.g., 18-predictor model) (Graham & Schafer, 1999).
4. Ignorability assumption is suspect (Graham et al., 1996, 1997).
5. Data distribution is skewed (Graham et al., 1996; Wang et al., 1999).
6. Model of the data distribution is misspecified (Greenland & Finkle, 1995).

Empirical and simulation studies have shown that MI is far superior to deletion procedures, mean substitution, regression imputation (Crawford et al., 1995; Graham et al., 1996), and simple hot deck procedure (DeCanio & Watkins, 1998) with regard to bias, efficiency, and validity of interval estimates when the underlying MI model specification is correct.

Limitations of Multiple Imputation

Since the observed scores Y_o provide indirect evidence about the likely values of the missing ones Y_m in MI, relevant predictors for Y (i.e., knowledge on the cause of missingness) are essential to obtain unbiased estimates and valid inferences.

Summary

A summary of the development, theoretical framework, and assumptions of MI were discussed. The procedures on performing MI based on a normal linear regression model with a univariate Y variable as well as how to combine k intermediate parameter estimates into a final adjusted point estimate and its variance were described. The two main features of MI are: (i) it takes into consideration the uncertainty of not knowing the exact values of the missing scores by incorporating the residual variation about the regression prediction, and (ii) it incorporates the sampling variability to estimate the population distribution of the missing scores, which are unknown. Because of these two features, MI has been shown to yield satisfactory parameter estimates with relatively little bias.

Missing Data Mechanism

Little and Rubin (1987) indicated that valid inferences from MI depend on the inclusion of a correct mechanism that produces missingness, and the knowledge of the missing data mechanism is important in selecting an appropriate imputation model. Rubin (1976) defined the mechanism of missingness in terms of a probability distribution model of nonresponse. Let R denote an $N \times 1$ vector of binary missing-data indicators with

distribution depending on a parameter vector ψ for the nonresponse model. If an examinee responds to an item, $R = 1$; if an examinee omits an item, $R = 0$.

Since $Y = (Y_o, Y_m)$, the probability distribution of the Y function for the complete data can be expressed as

$$f(Y | \theta) \equiv f(Y_o, Y_m | \theta) \quad (2-20)$$

Integrating over the sampling space of missing scores Y_m yields the marginal probability distribution for the observed scores (Schafer, 1997).

$$f(Y_o | \theta) = \int f(Y_o, Y_m | \theta) dY_m \quad (2-21)$$

The probability distribution for the observed scores given the parameters θ of the data model and the parameters ψ of the missing data mechanism can be expressed as

$$f(Y_o, R | \mathbf{X}, \theta, \psi) = \int f(Y_o, Y_m, R | \mathbf{X}, \theta, \psi) dY_m \quad (2-22)$$

where θ and ψ are sets of indexing vectors of unknown parameters for their respective distributions. For example, the parameters in ψ are the proportions of examinees assigned to each item (Bradlow & Thomas, 1998), whereas $\theta = (\mu, \sigma^2)$ or $\theta = (\beta, \sigma^2)$.

The equation (2-22) can be factorized

$$\text{as } f(Y_o, R | \mathbf{X}, \theta, \psi) = \int f(R | Y_o, Y_m, \mathbf{X}, \psi) f(Y_o, Y_m | \mathbf{X}, \theta) dY_m \quad (2-23)$$

where

$f(R | Y_o, Y_m, \mathbf{X}, \psi)$ denotes the conditional distribution of R given Y and represents a model for the missing data mechanism, and

$f(Y_o, Y_m | \mathbf{X}, \theta)$ represents a model for the data.

Little and Rubin (1987) distinguished three types of missing data mechanisms: missing completely at random (MCAR), missing at random (MAR), and nonignorable missing (NIM).

Missing Completely at Random

Missing data mechanism is MCAR if the probability of the missing data indicator, $f(R)$, is independent of both the observed scores Y_o and the missing scores Y_m in the model, which means

$$f(R | Y_o, Y_m, \mathbf{X}, \psi) = f(R | \psi) \text{ for all } Y \quad (2-24)$$

An example of MCAR in education is when the probability of missing responses to an item in an achievement test depends on neither the examinees' ability nor the number of instruction hours on test taking skills received, and the number of instruction hours received for all examinees are known.

As indicated in equation (2-24), the distribution of the missing data indicator R does not depend on the missing scores or any covariates \mathbf{X} . The first term of the marginal probability distribution in equation (2-23) can therefore come out of the integral, and equation (2-23) can be expressed as

$$f(Y_o, R | \mathbf{X}, \theta, \psi) = f(R | \psi) \int f(Y_o, Y_m | \theta) dY_m \quad (2-25)$$

From equation (2-21), when the MCAR assumption holds, the probability distribution of the observed scores then becomes

$$f(Y_o, R | \mathbf{X}, \theta, \psi) = f(R | \psi) f(Y_o | \theta) \quad (2-26)$$

where

$f(R | \psi)$ represents a model for the missing data mechanism, and

$f(Y_o | \theta)$ represents a model for the conditional probability distribution of the observed scores Y_o .

Since Y_o and R are independent in equation (2-26), the sampling distribution of the observed scores is a marginal of the complete data distribution (Laird, 1988). This situation implies that sampling-based inferences such as regression imputation that make use of the distributional properties of the marginal distribution of the observed scores are unbiased and valid (Heitjan, 1997). However, MCAR makes the strongest assumption among the three types of missing data mechanisms (Little, 1992).

The likelihood function of the observed scores under the MCAR assumption in equation (2-26) can be factorized into two components, one pertaining solely to the structural parameter θ of the model and the other pertaining solely to the nuisance parameter ψ of the missing data mechanism.

$$L(\theta, \psi | Y_o, X, R) \propto f(R | \psi) f(Y_o | \theta) \quad (2-27)$$

When the joint parameter space of (θ, ψ) is the product of the parameter space of each separately, that is the two parameters θ and ψ are independent, the likelihood of θ based on Y_o , $L(\theta | Y_o)$ is a function proportional to $f(Y_o | \theta)$ (Chiremba, 1995),

$$L(\theta | Y_o) \propto f(Y_o | \theta) \quad (2-28)$$

Since $L(\theta | Y_o)$ is proportional to $f(Y_o | \theta)$, it is not necessary to specify the missing data mechanism when using likelihood-based inferences to obtain unbiased estimates (Laird, 1988). Under the MCAR assumption, Bayesian inference or maximum likelihood estimation of the structural parameters θ will yield valid inferences from the

observed scores $f(Y_o | \theta)$ without estimating the parameters ψ (Rubin, 1976). That is why the missing data mechanism is ignorable for likelihood-based inferences.

The pattern of missingness on Y under the MCAR assumption is completely randomly determined. The MCAR assumption can be assessed by comparing distributions of missing variable Y for respondents and nonrespondents on covariates X to check evidence for a systematic difference between nonrespondents and respondents (Curran, Bacchi, Hsu Schmitz, Molenberghs, & Sylvester, 1998).

Missing at Random

MAR is based on a weaker assumption. Under the MAR assumption, the conditional probability distribution of missing data indicator R given X depends on the observed scores Y_o , but not the missing scores Y_m (Little, 1995)

$$f(R | Y_o, Y_m, X, \psi) = f(R | Y_o, X, \psi) \text{ for all } Y_m \quad (2-29)$$

For example, the probability of missing responses to an item in an achievement test depends on the scores of the measured variables (e.g., number of instruction hours on test taking skills received), but not on the missing scores of the item itself (e.g., item difficulty).

As indicated in equation (2-29), the distribution of the missing data indicator R does not depend on the missing scores. The first term of the marginal probability distribution in equation (2-23) can therefore come out of the integral, and equation (2-23) can be expressed as

$$f(Y_o, R | X, \theta, \psi) = f(R | Y_o, X, \psi) \int f(Y_o, Y_m | X, \theta) dY_m \quad (2-30)$$

As in equation (2-21), the probability distribution of Y_o is the marginal probability distribution.

$$f(Y_o | \mathbf{X}, \theta) = \int f(Y_o, Y_m | \mathbf{X}, \theta) dY_m \quad (2-31)$$

The probability distribution of the observed scores under the MAR assumption then becomes

$$f(Y_o, R | \mathbf{X}, \theta, \psi) = f(R | Y_o, \mathbf{X}, \psi) f(Y_o | \mathbf{X}, \theta) \quad (2-32)$$

When data are MAR, the sampling distribution of the observed scores no longer equals the ordinary marginal distribution, but depends upon the missing process (Laird, 1988). Hence sampling-based inferences are biased. On the other hand, the likelihood function of the observed scores under the MAR assumption in equation (2-32) can be factorized into two components:

$$L(\theta, \psi | Y_o, \mathbf{X}, R) \propto f(R | Y_o, \mathbf{X}, \psi) f(Y_o | \mathbf{X}, \theta) \quad (2-33)$$

Once again, when the two parameters θ and ψ are functionally unrelated, the likelihood of the structural parameters θ based on Y_o is a function proportion to the marginal probability distribution of Y_o (Rubin, 1976).

$$L(\theta | Y_o, \mathbf{X}) \propto f(Y_o | \mathbf{X}, \theta) \quad (2-34)$$

Thus the missing data mechanism under the MAR assumption is also ignorable for the likelihood-based inferences (Rubin, 1976). In summary, when the missing data mechanism is ignorable, the imputation model does not have to include a distribution of the missing data indicator R , and the likelihood function of θ is based on only the observed scores Y_o .

Rubin (1976) showed that the response mechanism generating the missing data is ignorable for likelihood-based inferences if the parameter θ of the data model and the parameter ψ associated with the missing data mechanism are independent or functionally unrelated, and the missing data are MAR.

Conceptually, the missing values under the MAR assumption are a random sample from the complete data after conditioning on the measured variables X in the imputation model; therefore, the process of creating these missing values can be modeled using these variables (Barnard, Du, Hill, & Rubin, 1998). For example, the percent of missing responses in an item of an achievement test differs in groups of examinees with high, medium, and low cognitive-ability scores, and the scores of the cognitive ability for all examinees are known. Under the MAR assumption, the missing responses are randomly distributed within these three subgroups of examinees, even when the responses are not missing at random across subgroups (Roth, 1994). In other words, the measured variables X (i.e., the cognitive ability in this example) can account for the differences in the distribution of Y between nonrespondents and respondents (Little & Schenker, 1995).

In addition to MAR, Roth (1994) identified another pattern of missingness when missing data are related to other variables. In this pattern, missing data are nonrandomly distributed across and within subgroups. For example, more scores are missing at the bottom range of the high cognitive-ability group but a relatively few missing scores at the top range of the same group (Roth, 1994).

Accessible Missing Data Mechanism

Graham and Donaldson (1993) defined the missing data mechanism as “accessible” when the cause of missingness has been measured, whereas “ignorable” refers to a combination of accessible and proper use of the cause of missingness for analysis. Graham, Hofer, and Piccinin (1994) explained that unless the cause of missingness is incorporated properly in the analysis, the mechanism will not be ignorable.

Schafer (1997) pointed out that whether the missing data mechanism is ignorable is closely related to the fullness of the observed scores Y_o , the relevant variables X (i.e., causes of missingness), and the complexity of the data model $f(Y_o | X, \theta)$. If Y_o and X contain a lot of information for predicting Y_m , and are incorporated properly in the imputation model for analysis, then the residual dependence of R upon Y_m after conditioning on Y_o and X will be small (Schafer, 1997). Including relevant variables X , (covariates, variables that relate to the nonresponse, and predictive variables that explain a considerable amount of variance of Y in the model) help to reduce the uncertainty of the imputations (van Buuren et al., 1999), and thus to adjust bias associated with the missing data (Graham et al., 1994, 1997).

Barnard and Meng (1999) advocated the adoption of a “sensible imputation model”, which incorporates as many relevant variables for the cause of missingness, and at the same time keeps the model-building and fitting feasible so as to reduce multicollinearity problems. It is suggested that including extra variables may affect precision, but not bias in the inferences (Rubin 1987); on the other hand, leaving out relevant causes of missingness will yield biased estimation (Schafer, 1999; Schafer & Olsen, 1998).

Nonignorable Missing

Under the NIM assumption, the conditional probability distribution of missing data indicator R given \mathbf{X} is a function of the missing scores Y_m , or the values of unmeasured relevant variables, and possibly also the observed scores Y_o (Laird, 1988). The unmeasured variables may be unavailable or inaccessible.

$$f(R | Y_o, Y_m, \mathbf{X}, \psi) \quad (2-35)$$

For example, the probability of missing responses to an item in an achievement test depends on the missing scores of the item itself (e.g., item difficulty) and/or the examinees' true unobserved parameter (e.g., examinees' ability).

Since the conditional probability distribution (2-35) can not be simplified, NIM involves joint probability distribution modeling of both the complete data $f(Y_o, Y_m | \mathbf{X}, \theta)$ for Y , the missing data mechanism $f(R | Y_o, Y_m, \mathbf{X}, \psi)$ and the joint estimation of θ and ψ from Y_o and R , respectively (Schafer, 1997).

Little and Rubin (1987) suggested two ways to factorize the joint distribution of the complete data Y and missing data indicator R . One is based on the selection models:

$$f(Y, R | \mathbf{X}, \psi) = f(Y | \theta) f(R | Y, \psi) \quad (2-36)$$

where

$f(Y | \theta)$ is the model for the complete data Y , and

$f(R | Y, \psi)$ is the model for the missing data mechanism.

The other is pattern-mixture models:

$$f(Y, R | \varphi, \pi) = f(Y | R, \varphi) f(R | \pi) \quad (2-37)$$

where

$f(Y | R, \varphi)$ represents the distribution of Y conditioning on the missing data indicator R , $f(R | \pi)$ represents the marginal distribution of the missing data indicator for whether or not Y is missing, and

φ and π are the two unknown parameters corresponding to the two distributions.

Selection models specify the precise form of the nonresponse model, whereas pattern-mixture models incorporate the assumption of the missing data mechanism through restrictions on the parameters (Little, 1995). When R is independent of Y (i.e., when $\theta = \varphi$ and $\psi = \pi$), the missing data mechanism becomes MCAR, and the selection models are equivalent to the pattern-mixture models.

The likelihood function on the observed scores under the NIM assumption include a missing data indicator R and the missing data parameters ψ .

$$L(\theta, \psi | Y_o, \mathbf{X}, R) \propto f(Y_o, R | \mathbf{X}, \theta, \psi) \quad (2-38)$$

The joint distribution for Y and R typically involves more parameters such as the YR interaction term than can be estimated from Y_o and R alone (i.e., under-identifiable) (Little, 1995). In order to make them identifiable so that valid likelihood-based inferences can be made about the marginal responses, a restriction on the assumption is required. Schafer (1997) suggested that a priori restrictions be imposed on either the joint parameter space for θ and ψ , or the Bayesian prior distribution (θ, ψ) .

Conceptually, under the NIM assumption, the distribution of respondents and nonrespondents on Y differs systematically, even after conditioning on the values of measured variables \mathbf{X} (Rubin & Schenker, 1991). Compared to equation (2-3), the posterior predictive probability distribution under the NIM assumption needs to include a

full specification of the probability model with the joint distribution of Y , the nonresponse pattern R , and the measured variables X .

$$f(Y_m | Y_o, X, R) = \int f(Y_m | Y_o, X, R, \theta) f(\theta | Y_o, X, R) d\theta \quad (2-39)$$

Sensitivity Analysis

Often time, little is known about the nonresponse mechanism that creates the missing responses in a particular achievement test. Missing responses can arise from a variety of reasons including a combination of ignorable and nonignorable mechanisms (Schafer & Olsen, 1988). However, distinguishing between ignorable and nonignorable mechanisms (i.e., MAR and NIM) relies on fundamentally untestable assumptions (Curran et al., 1998). Curran and associates (1998) demonstrated that these assumptions cannot be tested formally from the empirical data at hand. Analyses should be conducted to compare the estimates across a number of plausible missing-data models. Inferences from the sensitivity analysis reveal uncertainty about reasons for nonresponse (Beaton, 1997; Little & Rubin, 1989). Sensitivity analysis can also be conducted across alternative imputing procedures in a similar manner to reveal uncertainty about different possible imputation models.

Under the NIM assumption, sensitivity analysis can be performed by comparing estimates between selection and pattern-mixture models. If the results are consistent, confidence about the conclusions is established. On the other hand, if the results depend on the form of the model, then more specific conditions can be suggested about where the conclusion can apply (Little, 1995).

Summary

Valid inferences from MI rely on the inclusion of a correct missing data mechanism. As discussed above, factorization of the posterior predictive probability distribution depends on whether the missing data mechanism is ignorable or not. When the missing data mechanism is not ignorable, the missing data indicator has to be incorporated into the posterior predictive probability distribution, and the likelihood function of θ is not just based on the observed scores.

CHAPTER 3 METHODOLOGY

This chapter first describes the data generation procedure, the design of the study, and the MI procedure. Data generation was based on the three-parameter logistic model (Hambleton & Swaminathan, 1985). The design of this study involved the distribution and percent of missing responses as a function of the ability of the examinees and the difficulty of the items in one omitting pattern, and the ability of the examinees and the sequence of the items in another omitting pattern. The procedure of MI based on a logistic regression model with a univariate Y was outlined. The second part of this chapter discusses how to evaluate the effectiveness of MI on handling missing data.

To achieve the goal of evaluating the effectiveness of MI on handling missing data, several steps were required.

Step 1. Simulated a complete data matrix of item responses for a specified number of examinees.

Step 2. Computed the coefficient alpha. The coefficient alpha of this original complete data (i.e., 0% of missing) served as a benchmark for later comparison.

Step 3. Nonrandomly deleted certain percent of item responses from the complete examinee-by-item matrix generated in Step 1. Each missing data set was generated in a similar fashion.

Step 4. Replaced the omitting item responses in Step 3 using MI.

Step 5. Computed the coefficient alpha of the data set that was restored by MI.

Step 6. Compared the coefficient alpha from Step 2 with the one from Step 5.

Simulation Procedure

Let W be an $N \times P$ matrix representing a complete examinee-by-item data set. N is the number of examinees in the data set and P is the number of test items. In this study, P was fixed to be 20 in all conditions. A 20-item test was used because it represented test lengths frequently found in educational and psychological applications (Yen, 1987) such as the American Mathematical Association of Two-Year Colleges' Student Mathematics League contest (Isaacson & Smith, 1993). The response of each item was dichotomous in nature. Simulation of the 20 dichotomously-scored item responses of a specified number of examinees was based on the three-parameter logistic model (Hambleton & Swaminathan, 1985).

$$P_i(\xi_j) = c_i + (1 - c_i) \frac{\exp[Da_i(\xi_j - b_i)]}{1 + \exp[Da_i(\xi_j - b_i)]} \quad (3-1)$$

for $i = 1, \dots, 20$, and $j = 1, \dots, n$.

where

$P_i(\xi_j)$ is the probability of the j th examinee with an ability ξ_j answering the i th item correctly,

a_i is the discrimination parameter of item i ,

b_i is the difficulty parameter of item i ,

c_i is the pseudo-chance parameter of item i ,

ξ_j is the ability of the j th examinee, and

D is a scaling factor, which is -1.7 .

To compute $P(\xi)$, it required the three parameters (i.e., a , b , & c) and the ability parameters (i.e., ξ) to be known. The three parameters were drawn from the distributions of known mean and standard deviation. Harrison (1998) used the criteria of Oshima's (1994) study to sample the three parameters. The criteria for Oshima's (1994) study were as follows: the item discrimination parameters (i.e., a) were randomly drawn from a lognormal distribution with a mean of 1.13 and a standard deviation of 0.63; the item difficulty parameters (i.e., b) were randomly drawn from a normal distribution with a mean of 0 and a standard deviation of 1; and the pseudo-chance parameters (i.e., c) were randomly drawn from a normal distribution with a mean of 0.25 and a standard deviation of 0.05. According to Oshima (1994), the distributions of these three parameters were similar to the real data set of a speeded test (i.e., TOEFL) as reported by Way and Reese (1991). The ability parameters, ξ , for the examinees were randomly generated from a standard normal $N(0,1)$ distribution. The present study used the same values of the three item parameters as in Harrison's (1998, p. 7) study to generate the item responses (Table 3-1). The correlation between the item difficulty parameters (b) and the item discrimination parameters (a) was .111 ($p = .642$, two-tailed), whereas the correlation between the item difficulty parameters (b) and the pseudo-chance parameters (c) was .281 ($p = .230$, two-tailed).

The response data for a particular item given by an examinee with a trait level of ability, $P_i(\xi_j)$, was determined by computing the probability of correctly answering that item based on the known item and ability parameters. Since the item responses were dichotomous in nature, the response probabilities $P_i(\xi_j)$ were converted into binary responses by comparing to a random number r generated from a uniform distribution in

the interval between 0 and 1. The random number r was used to determine whether the score of a particular item was correct or incorrect. If the response probability obtained from the equation (3-1) was greater than the random number r , a 1 was assigned, which indicated the examinee's response to that particular item was correct; otherwise a 0 was assigned, which indicated the examinee's response to that particular item was incorrect.

Kuder-Richardson 20 Formula

Since the test items are scored dichotomously, Kuder-Richardson 20 formula (KR 20) was used to calculate the index of internal consistency for the test items. The Kuder-Richardson 20 formula is equivalent to coefficient alpha when the item responses are dichotomous in nature (Kuder & Richardson, 1937).

$$KR\ 20 = \frac{s}{s-1} \left[1 - \frac{\sum_{i=1}^s p_i q_i}{\sigma_x^2} \right] \quad (3-2)$$

where

s is the number of items in the test,

σ_x^2 is the variance of the test scores,

p_i is the proportion of subjects answering item correctly,

q_i is the proportion of subjects answering item incorrectly, and

$p_i q_i$ is the variance of scores on a single item i .

Table 3-1. Item Parameters Used for Test Simulation

Item	Discrimination (a)	Difficulty (b)	Pseudo-chance (c)
1	0.269	-0.772	0.176
2	0.236	-0.129	0.200
3	2.817	-0.979	0.257
4	8.565	0.235	0.193
5	1.452	0.072	0.264
6	1.043	-1.245	0.246
7	1.594	-1.504	0.229
8	1.258	0.545	0.221
9	5.502	-0.802	0.250
10	2.468	2.408	0.306
11	1.016	-0.048	0.231
12	3.413	2.062	0.240
13	2.238	0.262	0.287
14	2.370	-1.158	0.207
15	2.635	-0.314	0.276
16	0.533	-0.536	0.319
17	1.601	1.177	0.320
18	2.809	-0.471	0.261
19	0.036	-0.475	0.297
20	7.637	-0.203	0.328

Note. Adapted from Harrison (1998)

Design of Study

This study represents a $3 \times [(3 \times 3) + 1] \times 2$ design with three factors: sample size (3 levels), percent of examinees with missing items (3 levels), percent of items missing for each examinee with missing items (3 levels), and omitting pattern (2 levels) that were fully crossed. An additional condition with disproportional percent of examinees missing items that were nonrandomly distributed across and within each ability group was included. The rationales for selecting the levels in each factor were described below.

Sample Size

The three levels of sample size chosen in this study were: $N = 50, 100$, or 500 . The sample size of 50 examinees is typical for validation studies (Schmidt, Hunter, & Urry, 1976). The sample size of 500 was same as the real data that Raghunathana and Siscovick (1996) used in studying the performance of MI. These three levels, representing the small to large-sample sizes, were also used by Graham and Schafer (1999) to evaluate the efficiency of MI in a simulation study. The present study adopted these three levels of sample size to allow comparison of the performance of MI with that of other MDTs investigated by Harrison (1998).

Distribution and Percent of Missing Responses

In order to simulate a more realistic distribution and percent of nonresponse across test items, the distribution and percent of missing responses were based on the findings from a large scale study, the Reading Comprehension subtest, Level I, of the Comprehensive Test of Basic Skills, Form S (Cluxton & Mandeville, 1979). In Cluxton and Mandeville's (1979) study, they stratified one thousand third grade students into three ability levels--high, medium, and low. They found the proportion of students with

missing items within each stratified ability level was: 0-20% for the high ability group, 20-80% for the medium-ability group, and 90-100% for the low-ability group. They also reported the proportion of missing items (out of the 45 items in the subtest) for students within each stratified ability level was approximately: 7-16% for the high-ability group, 18-38% for the medium-ability group, and 40-49% for the low-ability group. The correlation between the ability of students and the number of items missing in the body of the test was $-.76$; and the correlation between the ability of students and the number of items missing at the end of the test was $-.47$ (Cluxton & Mandeville, 1979).

Based on the range of the proportion of students with missing items within each ability level, and the range of proportion of items missing provided in Cluxton and Mandeville's (1979) study, the distribution and percent of missing responses in this study were constructed in four steps:

First, the ability of the examinees in a sample were rank ordered. Second, the examinees in each data set were stratified into three ability levels. Stratification was based on the assumption that the data were normally distributed $N(0,1)$. Plus and minus one standard deviation in each sample were used as the cut-off to stratify the three ability groups. As a result, approximate 68% of the examinees were within the one standard deviation band and these students were classified as the medium-ability group. About 16% of the examinees were above one standard deviation and these students were classified as the high-ability group, and about 16% of examinees were below one standard deviation and these students were classified as the low-ability group.

For the percent of examinees with missing items (%EMI), three conditions (%EMI₁, %EMI₂, and %EMI₃) were constructed. In the first condition %EMI₁, the

percent of the high, medium, and low-ability examinees missing some test items were 0%, 20%, and 90% respectively. In the second condition %EMI₂, the percent of the high, medium, and low-ability examinees missing some test items were 10%, 50%, and 95%. In the third condition %EMI₃, the percent of the high, medium, and low-ability examinees missing some test items were 20%, 80%, and 100%. The above three conditions respectively corresponded to the minimum, the median, and the maximum percent of examinees with missing items in each ability level provided by Cluxton and Mandeville (1979).

Fourth, for the percent of items missing in those examinees with missing items responses (%IM), another three conditions (%IM₁, %IM₂, and %IM₃) were constructed. The first condition %IM₁ was 7% of the items missing in the high-ability group, 18% of the items missing in the medium-ability group, and 40% of the items missing in the low-ability group. The second condition %IM₂ was 12% of the items missing in the high-ability group, 28% of the items missing in the medium-ability group, and 45% of the items missing in the low-ability group. The third condition %IM₃ was 16% of the items missing in the high-ability group, 38% of the items missing in the medium-ability group, and 49% of the items missing in the low-ability group. The three conditions respectively corresponded to the minimum, the median, and maximum percent of items missing in each ability level provided by Cluxton and Mandeville (1979).

The above two sets of conditions were crossed to create nine missing conditions as shown in Figure 3-1. For example, the results of one combination were 20% of the high-ability examinees with three missing items (i.e., 16% of the 20 test items), 80% of the medium-ability examinees with eight missing items (i.e., 38% of the 20 test items),

and 100% of the low-ability examinees with ten missing items (i.e., 49% of the 20 test items). The distribution and percent of missing responses represented the typical range of missing data in educational tests, which is approximately 10-30% (Roth, 1994).

Range / Condition			Max / %IM ₁			Med / %IM ₂			Min / %IM ₃					
			Ability			H	M	L	H	M	L	H	M	L
			Percent	7	18	40	12	28	45	16	38	49		
Max / %EMI ₁	H	0												
	M	20												
	L	90												
Med / %EMI ₂	H	10												
	M	50												
	L	95												
Min / %EMI ₃	H	20												
	M	80												
	L	100												

Note. Max = Maximum, Med = Medium, and Min = Minimum.

Figure 3-1. Distribution and percent of missing responses in the nine missing conditions.

In addition to the nine missing conditions, an additional condition with disproportional percent of examinees omitted items that were nonrandomly distributed across and within each ability group was included (Roth, 1994). For example, more items were missing at the bottom range of the high-ability group but a relatively fewer items were missing at the top range of the same group (Roth, 1994). The procedure was to stratify each ability group (high, medium, and low) into three sub-strata. Stratification once again was based on plus and minus one standard deviation of the sample size within each of the three ability groups. The percent of examinees missed some items within each sub-stratum ability group (%EMI_s) was: 0, 10, 20 (in the high-ability group); 20, 50, 80 (in the medium-ability group); 90, 95, 100 (in the low-ability group). The corresponding percent of item missing within each ability sub-stratum (%IM_s) was: 7, 12, 16 (for the high-ability group); 18, 28, 38 (for the medium-ability group); and 40, 45, 49 (for the low-ability group). The two situations were then crossed to form the tenth condition. Table 3-2 summarized the distribution and percent of missing responses of the ten missing conditions.

Table 3-2. Summary the Distribution and Percent of Missing Responses of the Ten Missing Conditions

Condition	Description
$\%EMI_1 \times \%IM_1$	0% of the high-ability examinees having one missing item (i.e., 7% of the 20 test items) plus 20% of the medium-ability examinees having four missing items (i.e., 18% of the 20 items) plus 90% of the low-ability examinees having eight missing items (i.e., 40% of the 20 items). The total percent of missing data is approximately 8.4%.
$\%EMI_1 \times \%IM_2$	0% of the high-ability examinees having two missing items (i.e., 12% of the 20 test items) plus 20% of the medium-ability examinees having six missing items (i.e., 28% of the 20 items) plus 90% of the low-ability examinees having nine missing items (i.e., 45% of the 20 items). The total percent of missing data is approximately 10.5%.
$\%EMI_1 \times \%IM_3$	0% of the high-ability examinees having three missing items (i.e., 16% of the 20 test items) plus 20% of the medium-ability examinees having eight missing items (i.e., 38% of the 20 items) plus 90% of the low-ability examinees having ten missing items (i.e., 49% of the 20 items). The total percent of missing data is approximately 12.6%.
$\%EMI_2 \times \%IM_1$	10% of the high-ability examinees having one missing item (i.e., 7% of the 20 test items) plus 50% of the medium-ability examinees having four missing items (i.e., 18% of the 20 items) plus 95% of the low-ability examinees having eight missing items (i.e., 40% of the 20 items). The total percent of missing data is approximately 13.3%.

Condition	Description
$\%EMI_2 \times \%IM_2$	10% of the high-ability examinees having two missing items (i.e., 12% of the 20 test items) plus 50% of the medium-ability examinees having six missing items (i.e., 28% of the 20 items) plus 95% of the low-ability examinees having nine missing items (i.e., 45% of the 20 items). The total percent of missing data is approximately 17.6%.
$\%EMI_2 \times \%IM_3$	10% of the high-ability examinees having three missing items (i.e., 16% of the 20 test items) plus 50% of the medium-ability examinees having eight missing items (i.e., 38% of the 20 items) plus 95% of the low-ability examinees having ten missing items (i.e., 49% of the 20 items). The total percent of missing data is approximately 21.9%.
$\%EMI_3 \times \%IM_1$	20% of the high-ability examinees having one missing item (i.e., 7% of the 20 test items) plus 80% of the medium-ability examinees having four missing items (i.e., 18% of the 20 items) plus 100% of the low-ability examinees having eight missing items (i.e., 40% of the 20 items). The total percent of missing data is approximately 17.4%.
$\%EMI_3 \times \%IM_2$	20% of the high-ability examinees having two missing items (i.e., 12% of the 20 test items) plus 80% of the medium-ability examinees having six missing items (i.e., 28% of the 20 items) plus 100% of the low-ability examinees having nine missing items (i.e., 45% of the 20 items). The total percent of missing data is approximately 23.8%.
$\%EMI_3 \times \%IM_3$	20% of the high-ability examinees having three missing items (i.e., 16% of the 20 test items) plus 80% of the medium-ability examinees having eight missing items (i.e., 38% of the 20 items) plus 100% of the low-ability

Condition	Description
%EMI _s x %IM _s	<p>examinees having ten missing items (i.e., 49% of the 20 items). The total percent of missing data is approximately 30.2%.</p> <p>0% of the upper division of the high-ability examinees having one missing item (i.e., 7% of the 20 test items) plus 10% of the middle division of the high-ability examinees having two missing items (i.e., 12% of the 20 items) plus 20% of the lower division of the high-ability examinees having three missing items (i.e., 16% of the 20 items) plus 20% of the upper division of the medium-ability examinees having four missing items (i.e., 18% of the 20 test items) plus 50% of the middle division of the medium-ability examinees having six missing items (i.e., 28% of the 20 items) plus 80% of the lower division of the medium-ability examinees having eight missing items (i.e., 38% of the 20 items) plus 90% of the upper division of the low-ability examinees having eight missing items (i.e., 40% of the 20 test items) plus 95% of the middle division of the low-ability examinees having nine missing items (i.e., 45% of the 20 items) plus 100% of the lower division of the low-ability examinees having ten missing items (i.e., 49% of the 20 items). The total percent of missing data approximately 18.0%.</p>

Omitting Pattern

The two nonrandom omitting patterns were: (1) omitting item responses in the body of the test (OPB), and (2) omitting item responses at the end of the test (OPE) (i.e., non-reached). The first situation was similar to the missing data mechanism 5 in Harrison's (1998) study. However, in contrast to Harrison's (1998) study where examinees with the lowest abilities missed the most difficult items, examinees with different levels of abilities missed the most difficult items differentially. That meant that the high-ability examinees missed fewer difficult items than those of the medium-ability examinees, and in turn the medium-ability examinees missed fewer difficult items than those of the low-ability examinees (see Figure 3-2).

The non-reached pattern was similar to the combination of missing data mechanism 2 and 3 in Harrison's (1998) study. However, the selection of missing responses was based on the examinees' ability instead of random selection as the missing data mechanism 2 in Harrison's (1998) study. Once again, the high-ability examinees missed fewer items at the end of a test than those of the medium-ability examinees, and in turn the medium-ability examinees missed fewer items at the end of a test than those of the low-ability examinees (see Figure 3-3).

		Least difficult										Most difficult									
Person Ability	Item Difficulty	1	4	19	2	20	18	9	11	13	3	7	10	14	15	5	8	12	17	16	6
		1	1	0	1	0	1	0	0	0	1	1	0	0	1	1	1	1	1	1	
(high ability)	13	1	1	0	1	0	1	0	0	0	1	1	0	0	1	1	1	1	1	1	
	6	0	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	0	
	5	0	0	1	0	1	0	0	1	0	0	1	1	0	1	0	1	0			
	4	1	0	1	0	0	0	1	1	0	0	1	1	1	1	1	1				
	8	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	0	1			
	12	0	1	0	0	1	0	0	1	1	0	1	0	1	0	1					
	1	1	1	1	0	1	0	0	0	1	1	0	0	1	1	1					
	10	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1					
	13	1	0	0	1	0	0	1	0	0	1	1	0	1							
	3	1	0	1	1	1	1	1	1	1	0	1	1								
	2	1	1	1	1	0	1	0	1	1	0	0									
	15	0	1	0	0	1	1	0	1	0	0										
	14	1	1	0	0	1	1	0	0	0	1										
	7	1	1	0	0	1	0	0	0	1	1										
	9																				
(lower ability)		1	0	0	0	1	1	1	0	1	0										

Figure 3-2. Illustrate the omitting pattern of missing item responses in the body of the test (OPB) with 15 examinees and 20 test items.

Person Ability	Item																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
(high ability)	1	1	0	1	0	1	0	0	0	1	1	0	0	1	1	1	1	1	1	1
13																				
6	0	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	0	1	
5	0	0	1	0	1	0	0	1	0	0	1	1	0	1	0	1	0	1		
4	1	0	1	0	0	0	1	1	0	0	1	1	1	1	1	1	1			
8	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	0	1			
12	0	1	0	0	1	0	0	1	1	0	1	0	1	0	1					
1	1	1	1	0	1	0	0	0	1	1	0	0	1	1						
10	1	1	1	1	1	0	1	0	1	1	1	1	1	1						
13	1	0	0	1	0	0	1	0	0	1	1	0								
3	1	0	1	1	1	1	1	1	1	0	1									
2	1	1	1	1	0	1	0	1	1	0	0									
15	0	1	0	0	1	1	0	1	0	0										
14	1	1	0	0	1	1	0	0	0											
7	1	1	0	0	1	0	0	0	1											
9	1	0	0	0	1	1	1	0	1											
(lower ability)																				

Figure 3-3. Illustrate the omitting pattern of missing item responses at the end of the test (OPE) with 15 examinees and 20 test items.

Cause of Missingness

Under the omitting pattern in which item responses was omitted in the body of the test, the examinees' ability and the item difficulty provided an indirect evidence about the likely values of the missing responses. On the other hand, when the item responses were omitted at the end of the test, the examinee's ability and the item effect, which is the random effect of the test items, provide an indirect evidence about the likely values of the missing responses. Since the cause of missingness in this study was under the researcher's control, and the differential amount of missing responses was a function of the examinees' ability and item difficulty, or a function of the examinee's ability and item effect depending on the omitting patterns, the missing data mechanism could be considered as missing at random (Graham et al., 1997). The missing data mechanism was therefore ignorable.

Iterations

For each of the 60 conditions (3 levels of sample size, 10 levels of the distribution and percent of item response omission, and 2 levels of omitting pattern), one thousand iterations were performed to ensure stable results. The one thousand iterations have been used in two previous simulation studies in the evaluation of the efficiency of MI (Glynn, Laird, & Rubin, 1993; Graham et al., 1996). The iterations resulted in generating 1,000 repeated data set for each level of sample size.

Multiple Imputation Procedure

Let Y be an $N \times 1$ vector of measures with $Y \sim N(X\beta, \sigma^2)$, where $X\beta$ covariate was a function of the parameters, X was a matrix of examinees' ability and item difficulty

variables under the situation when the omitting item responses were in the body of the test, or examinees' ability and item effect variables under the situation when the omitting item responses were at the end of the test, and β was a vector of regression parameters to be estimated. The distribution of β was assumed to be multivariate normal. The algorithm for creating ten multiply imputed Y_m involved the following steps (Freedman, 1990; Freedman & Wolf, 1995):

Step 1. Specified a particular form of imputation model to predict the value of a missing variable Y and estimate the parameter vector β of regression coefficients using the portion of the sample with complete data. Since the item responses in this study were dichotomous in nature, the prediction model required a logistic regression model with a univariate Y of the form

$$\text{Logit}(p_j) = \ln\left(\frac{p_j}{1-p_j}\right) = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} \quad (3-3)$$

where $j = 1, \dots, n$ examinees.

The set of predictors entered the explicit model to create imputations for OPB differed from that for OPE. Under the situation when the omitting item responses were in the body of the test, X_{1j} was the examinee's ability, and X_{2j} was the item difficulty, whereas under the situation when the omitting response were at the end of the test, X_{1j} was the examinee's ability, and X_{2j} was the item effect, which was the random effect of the 20 items. As a result, there were two distinct logistic regression models for the MI procedure.

The posterior probability of the response given X_{1j} and X_{2j} was

$$p_j = E(Y_j | X_1, X_2) = \Pr(Y_j = 1 | X_{1j}, X_{2j}) \quad (3-4)$$

The logistic regression model assumed that the logit of the posterior probability was a linear combination of the X_{1j} and X_{2j} variables.

$$Y_j = \begin{cases} 1 & \text{if } \text{logit} [\hat{\beta}' X_j] > \hat{u}_j \\ 0 & \text{otherwise,} \end{cases} \quad (3-5)$$

where \hat{u}_j was a random error.

Regressed Y_o on the corresponding X_o matrix gave the ordinary least squares estimates: estimated regression coefficient vector $\hat{\beta}$ and estimated variance-covariance matrix Σ .

Step 2. Randomly drew from the sampling distribution of regression coefficients.

Estimated the regression coefficients by adding the random error to account for the uncertainty about the regression prediction, which was the same as equation (2-11). Each repetition used a distinct value of β^* common across all imputed cases. This value was drawn independently from the multivariate normal distribution of the estimated vector $\hat{\beta}$.

Step 3. Given an estimate of β and the value for X_1, X_2 , the probability of answering an item correctly could be predicted with equation (3-5) by drawing a value of u^* drawn from the uniform $[0,1]$ distribution. Each set of predicted values Y_m^* was based on different sets of regression predictions and an independent drawn value of u^* . The probability of a correct response for respondent j in the k th repetitions, $p_{j(k)}$, was calculated with the randomly selected regression coefficients and the value of j for the corresponding covariates from the logistic regression:

$$p_{j(k)} = \frac{e^{\beta_{(1)}X_{1j} + \beta_{(2)}X_{2j}}}{1 + e^{\beta_{(1)}X_{1j} + \beta_{(2)}X_{2j}}}$$

Step 4. The estimated probability $p_{j(k)}$ from the logistic regression was compared to a random number t from the uniform $[0,1]$ distribution for each missing score. If the predicted probability $p_{j(k)}$ was less than t then the imputed value for $Y_{j(k)}$ was assigned a 1; otherwise the imputed value was assigned a 0. These probabilities were used to impute the missing scores.

Step 5. Conducted ten repetitions which meant repeating steps 2 to 4 ten times to create a series of ten imputed values (i.e., ten distinct imputed data sets).

Step 6. Computed the KR 20 (i.e., coefficient alpha) separately in each of the ten imputed complete data sets. This resulted in ten separate coefficient alphas.

Step 7. Using the equation (2-15), the final adjusted coefficient alpha was obtained by taking the simple arithmetic average of the ten coefficient alphas. This final coefficient alpha was then compared to the one obtained from the original complete data set.

Evaluating the Performance of Multiple Imputation

The accuracy (bias and precision) of the coefficient alpha obtained from the restored complete data set in each of the ten missing conditions using MI was assessed by means of the bias and the root-mean-square error (RMSE). Measures of the bias and RMSE were averaged over the 1,000 iterations of the simulation.

Bias is defined as the average value of the coefficient alphas derived from the original complete data set with no missing data minus the average value of the coefficient alphas from the corresponding imputed data set over the 20,000 (i.e., $2 \times 10 \times 1000$) completed tests for a particular number of examinees. The estimated coefficient alpha is

unbiased when the average deviation (i.e., bias) between the coefficient alpha obtained from the imputed values and that of the original values in the data set is close to 0.

RMSE is defined as the square root of the average squared difference between the coefficient alpha derived from the original complete data set with no missing data and the coefficient alpha from the corresponding imputed data set. The estimated coefficient alpha is precise when the RMSE is close to 0.

$$RMSE = \sqrt{\sum (\alpha \text{ of original data} - \alpha \text{ of restored data using MI})^2} \quad (3-6)$$

The relationship between RMSE and bias is

$$(RMSE)^2 = (\text{bias})^2 + (SE)^2 \quad (3-7)$$

where

RMSE is the root mean square error, which represents an overall error,

bias is the average deviation, which represents a systematic error, and

SE is the standard error, which represents random error.

CHAPTER 4 RESULTS

In this chapter results of the analyses of the data for the two performance criteria are presented. The two criteria are the bias and root mean square error (RMSE). The mean coefficient alpha and its standard deviation of the original complete data set with no missing data for 50, 100, and 500 examinees were $M = 0.765$, $SD = 0.033$; $M = 0.764$, $SD = 0.023$; $M = 0.763$, $SD = 0.01$ respectively. Each mean coefficient alpha was based on the average of 20,000 (10 missing conditions x 2 omitting patterns x 1000 iterations) values. The results of these mean coefficient alphas were very close to those computed by Harrison (1998). For example, the mean coefficient alpha for the sample size of 50 in Harrison's study was 0.769. The mean coefficient alphas and their standard deviation of the restored completed data set using MI for the ten missing conditions in each of the three levels of sample size and two levels of omitting pattern are shown in Figures 4-1 to 4-6.

The biases obtained in each of the ten missing conditions for the two omitting patterns are summarized in Tables 4-1 and 4-2. Under the omitting pattern where missing responses are at the end of the test (OPE), the biases (in absolute value) ranged from 0.000 to 0.030. The majority (93%) of the biases in OPE were in the magnitude of less than 0.02. The biases (in absolute value) obtained in OPE for the sample size 50, 100, and 500 ranged from 0.001 to 0.030, 0.000 to 0.016, and 0.000 to 0.009 respectively. The biases obtained in the omitting pattern where missing responses are in the body of the

test (OPB) were noticeably higher than those in OPE of the corresponding missing conditions, the), the biases (in absolute value) ranged from 0.019 to 0.069. The majority (97%) of the biases in OPB were less than 0.06. The biases (in absolute value) obtained in OPB for the sample size 50, 100, and 500 ranged from 0.027 to 0.069, 0.019 to 0.051, and 0.028 to 0.054 respectively. As expected, the largest bias was in the missing condition $\%EMI_3 \times \%IM_3$ where the small sample size (50) accompanied with the largest percent (30.2%) of missingness. The bias in this condition was 0.069. The coefficient alphas in OPB were always overestimated (positively biased), whereas in OPE, about half of the coefficient alphas obtained through MI was overestimated and the other half was underestimated. Whether MI produced the coefficient alphas that were overestimated or underestimated in OPE did not depend on the percent of missingness. Further research needs to be conducted to explore why some of the coefficient alphas obtained from MI were overestimated while others were underestimated under the same omitting pattern.

For condition $\%EMI_4 \times \%IM_4$ in which nonrandom distribution of omission is across and within each ability group (Roth, 1994), the bias obtained in this condition, regardless of the omitting patterns, was similar to other missing conditions where the percent of missingness was about the same (e.g., missing condition $\%EMI_2 \times \%IM_2$). The RMSEs obtained in each of the ten missing conditions for the two omitting patterns are summarized in Tables 4-3 and 4-4. The results were very similar to those obtained for the bias.

In general, the bias (in absolute value) or the RMSE increased as the amount of missingness increased. Graham and Schafer (1999) explained this phenomenon by suggesting that MI introduced bias when handling missing data. However, the pattern of

increment in the bias or the RMSE was not unidirectional as indicated in Tables 4-1 to 4-4. There were irregularities in the magnitude of the bias or the RMSE across the ten missing conditions within each sample size. That means in some missing conditions, the magnitude of the bias or the RMSE for the smaller amount of missingness was bigger than that of the larger amount of missingness even both conditions had the same sample size. This kind of irregularity was also noticed in Graham and Schafer's (1999) simulation study. Another general pattern revealed in this study was that the bias decreased as the sample size increased. Once again the pattern of increment in the bias was not unidirectional as indicated in Tables 4-1 and 4-2. There were irregularities across the three sample sizes, and this kind of irregularity was also noticed in Graham and Schafer's study. On the other hand, the magnitude of the RMSE in OPE, but not in the OPB showed a clear pattern of decrement as the sample size increased (see Table 4-3 and 4-4).

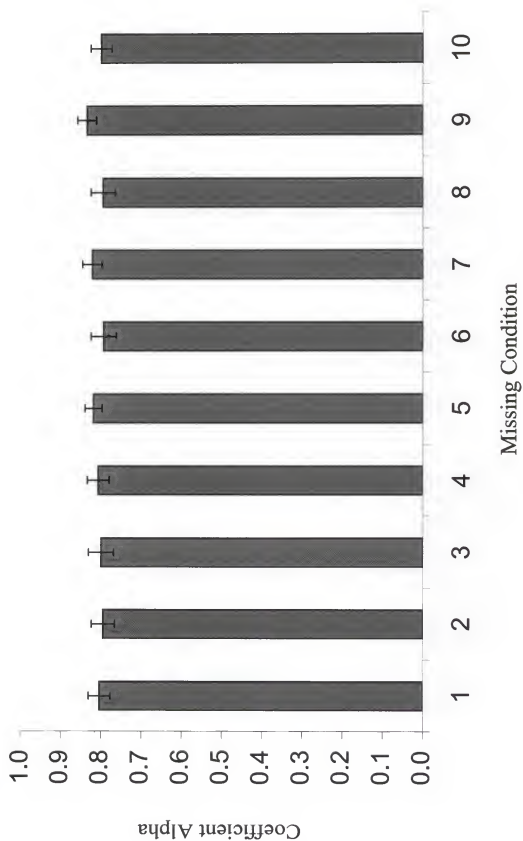


Figure 4-1. The mean coefficient alpha in OPB with sample size of 50.

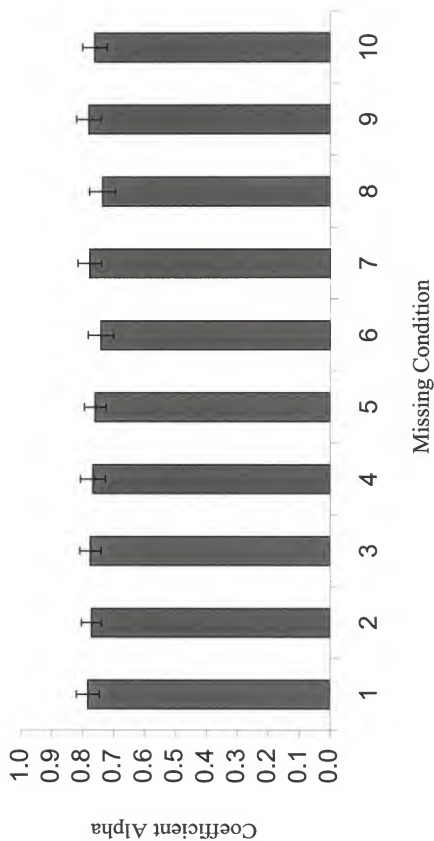


Figure 4-2. The mean coefficient alpha in OPE with sample size of 50.

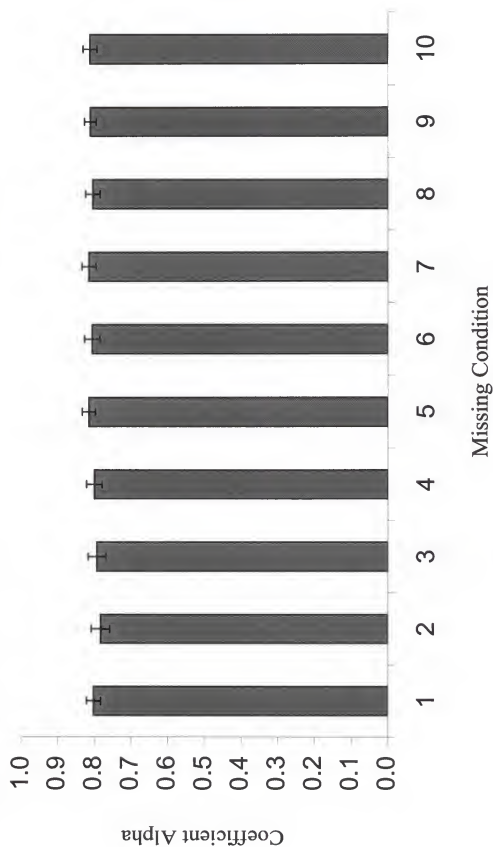


Figure 4-3. The mean coefficient alpha in OPB with sample size of 100.

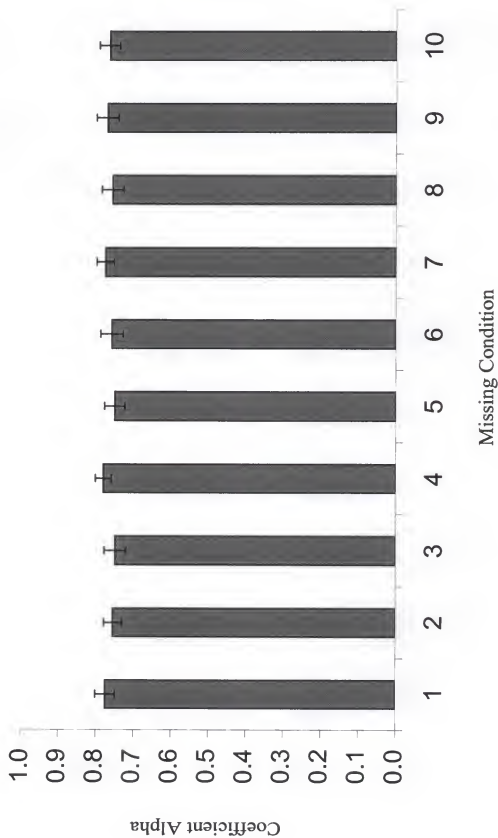


Figure 4-4. The mean coefficient alpha in OPE with sample size of 100

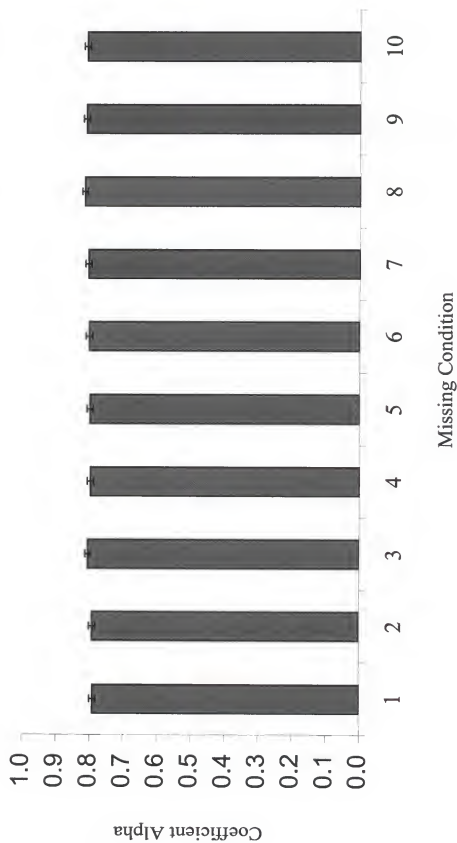


Figure 4-5. The mean coefficient alpha in OPB with sample size of 500.

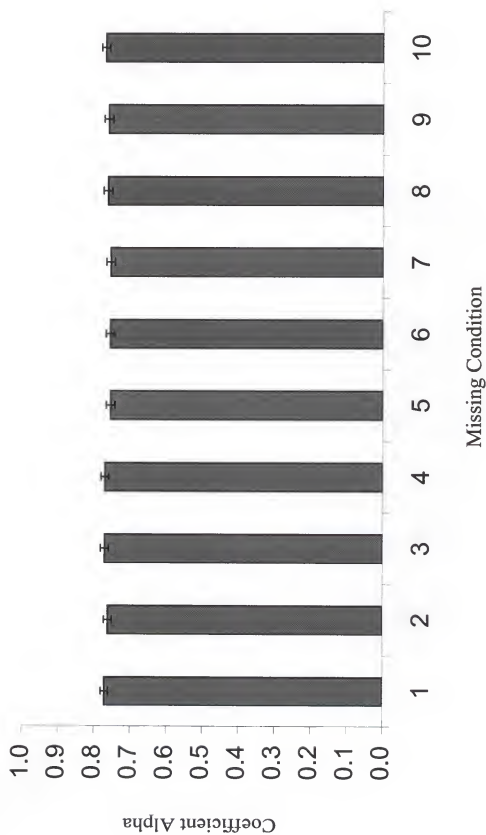


Figure 4-6. The mean coefficient alpha in OPE with sample size of 500.

Table 4-1. Bias for the Coefficient Alpha in Omitting Pattern Where Missing Responses are in the Body of the Test

		Sample size		
		50	100	500
Missing Condition	Approx. % of Missingness	Mean (SD)	Mean (SD)	Mean (SD)
%EMI ₁ x %IM ₁	8.4%	0.036 (0.017)	0.037 (0.012)	0.028 (0.005)
%EMI ₁ x %IM ₂	10.5%	0.028 (0.018)	0.019 (0.014)	0.029 (0.006)
%EMI ₁ x %IM ₃	12.6%	0.035 (0.019)	0.030 (0.015)	0.043 (0.006)
%EMI ₂ x %IM ₁	13.3%	0.042 (0.020)	0.036 (0.013)	0.036 (0.006)
%EMI ₃ x %IM ₁	17.4%	0.059 (0.021)	0.051 (0.015)	0.043 (0.007)
%EMI ₂ x %IM ₂	17.6%	0.054 (0.022)	0.050 (0.015)	0.035 (0.007)
%EMI ₁ x %IM ₄	18.0%	0.034 (0.020)	0.047 (0.015)	0.047 (0.007)
%EMI ₂ x %IM ₃	21.9%	0.027 (0.023)	0.043 (0.016)	0.040 (0.007)
%EMI ₃ x %IM ₂	23.8%	0.031 (0.024)	0.040 (0.016)	0.054 (0.007)
%EMI ₃ x %IM ₃	30.2%	0.069 (0.023)	0.047 (0.018)	0.048 (0.008)

Table 4-2. Bias for the Coefficient Alpha in Omitting Pattern Where Missing Responses are at the End of the Test

		Sample size		
		50	100	500
Missing Condition	Approx. % of Missingness	Mean (SD)	Mean (SD)	Mean (SD)
%EMI ₁ x %IM ₁	8.4%	0.017 (0.019)	0.011 (0.014)	0.007 (0.006)
%EMI ₁ x %IM ₂	10.5%	0.006 (0.017)	-0.010 (0.014)	-0.002 (0.006)
%EMI ₁ x %IM ₃	12.6%	0.009 (0.020)	-0.015 (0.017)	0.007 (0.007)
%EMI ₂ x %IM ₁	13.3%	0.001 (0.023)	0.016 (0.013)	0.007 (0.006)
%EMI ₃ x %IM ₁	17.4%	0.013 (0.024)	0.010 (0.014)	-0.009 (0.008)
%EMI ₂ x %IM ₂	17.6%	-0.004 (0.022)	-0.013 (0.016)	-0.008 (0.007)
%EMI ₄ x %IM ₄	18.0%	-0.005 (0.024)	0.000 (0.018)	0.005 (0.007)
%EMI ₂ x %IM ₃	21.9%	-0.024 (0.025)	-0.006 (0.020)	-0.008 (0.008)
%EMI ₃ x %IM ₂	23.8%	-0.030 (0.028)	-0.008 (0.020)	0.000 (0.008)
%EMI ₃ x %IM ₃	30.2%	0.017 (0.029)	0.006 (0.021)	-0.003 (0.009)

Table 4-3. RMSE for the Coefficient Alpha in Omitting Pattern Where Missing Responses are in the Body of the Test

Missing Condition	Approx. % of Missingness	Sample size		
		50	100	500
		Mean (SD)	Mean (SD)	Mean (SD)
%EMI ₁ x %IM ₁	8.4%	0.037 (0.017)	0.037 (0.012)	0.028 (0.005)
%EMI ₁ x %IM ₂	10.5%	0.029 (0.017)	0.020 (0.013)	0.029 (0.006)
%EMI ₁ x %IM ₃	12.6%	0.035 (0.018)	0.030 (0.014)	0.043 (0.006)
%EMI ₂ x %IM ₁	13.3%	0.043 (0.019)	0.036 (0.013)	0.036 (0.006)
%EMI ₃ x %IM ₁	17.4%	0.059 (0.021)	0.051 (0.015)	0.043 (0.007)
%EMI ₂ x %IM ₂	17.6%	0.054 (0.022)	0.050 (0.015)	0.035 (0.007)
%EMI ₁ x %IM ₄	18.0%	0.034 (0.019)	0.047 (0.015)	0.047 (0.007)
%EMI ₂ x %IM ₃	21.9%	0.030 (0.019)	0.043 (0.016)	0.040 (0.007)
%EMI ₃ x %IM ₂	23.8%	0.033 (0.021)	0.040 (0.016)	0.054 (0.007)
%EMI ₃ x %IM ₃	30.2%	0.069 (0.023)	0.047 (0.018)	0.048 (0.008)

Table 4-4. RMSE for the Coefficient Alpha in Omitting Pattern Where Missing Responses are at the End of the Test

Missing Condition	Approx. % of Missingness	Sample size		
		50	100	500
		Mean (SD)	Mean (SD)	Mean (SD)
%EMI ₁ x %IM ₁	8.4%	0.021 (0.014)	0.015 (0.010)	0.008 (0.005)
%EMI ₁ x %IM ₂	10.5%	0.015 (0.011)	0.014 (0.010)	0.005 (0.004)
%EMI ₁ x %IM ₃	12.6%	0.018 (0.014)	0.018 (0.013)	0.008 (0.005)
%EMI ₂ x %IM ₁	13.3%	0.018 (0.014)	0.017 (0.011)	0.008 (0.005)
%EMI ₃ x %IM ₁	17.4%	0.022 (0.016)	0.014 (0.011)	0.010 (0.006)
%EMI ₂ x %IM ₂	17.6%	0.018 (0.014)	0.017 (0.013)	0.009 (0.006)
%EMI ₄ x %IM ₄	18.0%	0.020 (0.015)	0.014 (0.011)	0.007 (0.005)
%EMI ₂ x %IM ₃	21.9%	0.028 (0.021)	0.016 (0.013)	0.009 (0.006)
%EMI ₃ x %IM ₂	23.8%	0.033 (0.024)	0.017 (0.013)	0.006 (0.005)
%EMI ₃ x %IM ₃	30.2%	0.027 (0.020)	0.017 (0.013)	0.007 (0.006)

CHAPTER 5 DISCUSSION

Because there was no substantial bias for all the missing conditions, the results of this simulation study indicated that MI is a reasonably good procedure to replace the missing data in a single-facet crossed model in which missing responses are either in the body of the test or at the end of the test. The majority of the biases obtained were less than 0.05, and the magnitude was comparable to those obtained in Harrison's (1998) study. The most significant difference was that the amount of missingness in the present study was two to three times more than that used in Harrison's study, and the omitting patterns were nonignorable.

The present study used the examinee's ability ξ and item difficulty b as the predictors in the logistic regression when the missing responses were in the body of the test, and the examinee's ability ξ and item effect i as the predictors when the missing responses were at the end of the test. The predictors used in the present study differed from the ones used by Harrison (1998). Harrison (1998) used examinee effect j and item effect i as the predictors. Results of using j and i as the predictors in the present study indicated that the biases for the coefficient alpha were unacceptably higher than those obtained using ξ and b , or ξ and i . For example, in the missing condition %EMI₃ x %IM₃ with a sample size of 50, the bias obtained using the j and i as the predictors was -0.211 when missing responses were in the body of the test, and -0.233 when missing responses were at the end of the test. This illustrated one of the limitations of MI as mentioned in

Chapter 3, namely that inference based on MI will be biased when relevant predictors are not incorporated (Schafer, 1999; Schafer & Olsen, 1998).

An attempt to include more predictors (i.e., examinee's ability ξ , item difficulty b , examinee effect j and item effect i) in the logistic model did not help to reduce the bias. For example, the bias obtained using ξ , b , j and i as the predictors was 0.07 in the missing condition %EMI₃ x %IM₃ when the omitting pattern was OPB and the sample size was 50. This illustration affirmed Rubin's (1987) suggestion that extra variables did not affect the bias in the inferences. Further systematic studies need to be conducted to support Rubin's claim regarding the relationship between the bias and the number of predictors.

Another possible factor that may affect the accuracy of the obtained coefficient alphas was the extreme value in some of the item discrimination parameters (e.g., $a = 7.637$). Unfortunately, a simpler model such as a one-parameter Rasch model with fixed item discrimination and pseudo-chance parameters did not help to reduce the bias. The bias for the above missing condition (i.e., %EMI₃ x %IM₃) was still in the magnitude of 0.07 when using ξ and b as the predictors.

When comparing the biases obtained from the two omitting patterns, it is suggested that examinee's ability rather than item effect or item parameters may contribute more to the accuracy of the parameter estimation. Further systematic investigation is warranted.

Finally, a surprising finding was obtained when using listwise deletion to estimate the coefficient alpha in the above missing condition (i.e., %EMI₃ x %IM₃)—the bias was -0.077. The bias (in absolute value) was much smaller than those obtained from Harrison's study (1998), even the amount of missingness was three time more. The bias

obtained from the nonrandom missing conditions (with 10% of missingness) in Harrison's study was about -0.2. This surprising finding may have something to do with the idiosyncratic nature of the missing mechanism in this study. Further research need to systematic investigate this issue.

Limitations

The present study used examinee's ability and item difficulty as the predictors in OPB, and used examinee's ability and item effect as the predictors in OPE. However, in real life testing situations, the ability parameter ξ and item parameter b need to be estimated first. Accurate estimation of these two parameters may not be possible in situations with a substantial amount of missing data (Bradlow & Thomas, 1998). Another limitation is that one may not be sure of the mechanism producing the missing data.

Suggestions for Future Research

The present study only illustrated one way of using MI to analyze the data. It is important to perform a sensitivity analysis to compare the results obtained in the present study with those when the nonresponse model was treated as nonignorable. A comparison of the coefficient alpha obtained using the selection approach model versus the pattern-mixture model certainly would be informative.

The bias obtained in the present study as well as in Graham and Schafer's (1999) study was not a linear function of the amount of missingness or the sample size. However, no good explanation can be given based on the limited information provided in

the present study as well as in Graham and Schafer's (1999) study. This may be an important issue for further investigation.

Because of the positively skewed distribution of the biases in OPB and the lower bound nature of the coefficient alpha, it is suggested that using the median instead of the mean to compute the final adjusted alpha may be worthwhile to investigate.

This study illustrated two of the most commonly encountered omitting patterns—missing responses in the body of the test and at the end of the test. In most real life educational tests, the types of omitting patterns are much more complicated and the missing responses as suggested by Schafer and Olsen (1988) can arise from a variety of reasons including a combination of ignorable and nonignorable mechanisms. Systematic investigation of the effectiveness of different MDTs especially RMLE and MI under the conditions involving a combination of ignorable and nonignorable is important for examining different kinds of missing responses.

In chapter 2, several methods have been described to create the posterior predictive probability distribution from Y_o ; however, today few studies have attempted to compare different methods of data simulation applied to MI. Duncan, Duncan, and Li (1998) illustrated the use of data augmentation and bootstrap in a structural equation model. More studies should investigate the effectiveness of these data simulation methods.

Obviously, the application of MI is not confined to the single-facet situation; further study should explore the application of MI to multi-facet situations where a generalizability coefficient is obtained. The incorporation of rater facet in nested designs

can be an extension of the present study to test the effectiveness of MI in handling missing data in a more complicated situation.

REFERENCES

Angoff, W. H., & Schrader, W. B. (1984). A study of hypotheses basic to the use of rights and formula scores. Journal of Educational Measurement, 21, 1-17.

Bacik, J. M., Murphy, S. A., & Anthony, J. C. (1998). Drug use prevention data, missed assessments and survival analysis. Multivariate Behavioral Research, 33, 573-588.

Barnard, J., Du, J. T., Hill, J. L., & Rubin, D. B. (1998). A broader template for analyzing broken randomized experiments. Sociological Methods and Research, 27, 285-317.

Barnard, J., & Meng, X. L. (1999). Applications of multiple imputation in medical studies: From AIDS to NHANES. Statistical Methods in Medical Research, 8, 17-36.

Beaton, A. E. (1997). Missing scores in survey research. In J. P. Keeves (ed.), Educational research, methodology, and measurement: An international handbook (2nd ed., pp. 763-766). New York: Pergamon Press.

Bradlow, E. T., & Thomas, N. (1998). Item response theory models applied to data allowing examinee choice. Journal of Educational and Behavioral Statistics, 23, 236-243.

Brownstone, D., & Valletta, R.G. (1996). Modeling earnings measurement error: A multiple imputation approach. Review of Economics and Statistics, 78, 705-717.

Chiremba, A. M. (1995). Direct versus indirect methods for the estimation of variance-covariance matrices and regression parameters when data are skewed and incomplete. Unpublished doctoral dissertation, University of Florida, Gainesville.

Cluxton, S. E., & Mandeville, G. K. (1979, April). Latent trait models: Ability estimates and omitted items. Paper presented at the 63rd Annual Meeting of the American Educational Research Association, San Francisco, CA.

Crawford, S. L., Tennstedt, S. L., & McKinlay, J. B. (1995). A comparison of analytic methods for non-random missingness of outcome data. Journal of Clinical Epidemiology, 48, 209-219.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart, & Winston.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, DC: American Council on Education.

Curran, D., Bacchi, M., Hsu Schmitz, S. F., Molenberghs, G., & Sylvester, R. J. (1998). Identifying the types of missingness in quality of life data from clinical trials. Statistics in Medicine, 17, 739-756.

DeCanio, S.J., & Watkins, W.E. (1998). Investment in energy efficiency: Do the characteristics of firms matter? Review of Economics and Statistics, 80, 95-107.

Downey, D. G., & King, C. V. (1998). Missing data in Likert ratings: A comparison of replacement methods. Journal of General Psychology, 125, 175-191.

Duncan, T. E., Duncan, S. C., & Li, F. (1998). A comparison of model- and multiple imputation-based approaches to longitudinal analyses with partial missingness. Structural Equation Modeling, 5, 1-21.

Freedman, V. (1990). Using SAS to perform multiple imputation. (Discussion Paper Series UI-PSC-6). Washington, DC: Urban Institute,

Freedman, V., & Wolf, D. A. (1995). A case-study on the use of multiple imputation. Demography, 32, 459-470.

Gelfand, A. E., & Smith, A. M. F. (1990). Sampling based approaches to calculating marginal densities. Journal of the American Statistical Association, 86, 398-409.

Glynn, R. J., Laird, N. M., & Rubin D. B. (1993). Multiple imputation in mixture models for nonignorable nonresponse with follow-ups. Journal of the American Statistical Association, 88, 984-993.

Graham, J. W., & Donaldson, S. I. (1993). Evaluating interventions with differential attrition: The importance of nonresponse mechanisms and use of follow-up data, Journal of Applied Psychology, 78, 119-128.

Graham, J. W., Hofer, S. M., Donaldson, S. I., MacKinnon, D. P., & Schafer, J. L. (1997). Analysis with missing data in prevention research. In K. J. Bryant, M. Windle, & S. G. West (Eds.), The science of prevention: Methodological advances from alcohol and substance abuse research (pp. 325-366). Washington, DC: American Psychological Association.

Graham, J. W., Hofer, S. M., & McKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planning missing value patterns: An application of maximum likelihood procedures. Multivariate Behavioral Research, 31, 197-218.

Graham, J.W., Hofer, S.M., & Piccinin, A.M. (1994). Analysis with missing data in drug prevention research. In L. M. Collins & L. A. Seitz (Eds.), Advances in data analysis for prevention intervention research (NIDA Research Monograph 142, pp. 13-62). Washington, DC: National Institute on Drug Abuse.

Graham, J. W., & Schafer, J. L. (1999). On the performance of multiple imputation for multivariate data with small sample size. In R. H. Hoyle (Ed.), Statistical strategies for small sample research (pp. 1-29). Thousand Oaks, CA: Sage.

Greenland, S., & Finkle, W. D. (1995). A critical look at methods for handling missing covariates in epidemiologic regression analyses. American Journal of Epidemiology, 142, 1255-1264.

Gross, A. L. (1997). Interval estimation of bivariate correlations with missing data on both variables: A Bayesian approach. Journal of Educational and Behavioral Statistics, 22, 407-424.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer-Nijhoff.

Harrison, J. M. (1998). A comparison of strategies for estimating internal consistency on tests with missing data. Unpublished master's thesis, University of Florida, Gainesville.

Heitjan, D. F. (1997). Annotation: What can be done about missing data? Approaches to imputation. American Journal of Public Health, 87, 548-550.

Heitjan, D. F., & Little, R. J. A.. (1991). Multiple imputation for the fetal accident report and system. Applied Statistics, 40, 13-29.

Heitjan, D. F., & Rubin, D. B. (1990). Inference from coarse data via multiple imputation with application to age heaping. Journal of the American Statistical Association, 85, 304-314.

Isaacson, J., & Smith, G. (1993). Hosting a mathematics tournament for two-year college students. (ERIC Document Reproduction Service No. ED 366 382)

Jamshidian, M., & Bentler, P. M. (1999). ML estimation of mean and covariance structures with missing data using complete data routines. Journal of Educational and Behavioral Statistics, 24, 21-41.

Kalton, G., & Kasprzyk, D. (1986). The treatment of missing survey data. Survey Methodology, 12, 1-16.

Kim, J. O., & Curry, J. (1977). The treatment of missing data in multivariate analysis. Sociological Methods and Research, 6, 215-241.

Koretz, D., Lewis, E., Skewes-Cox, T., & Burstein, L. (1993). Omitted and non-reached items in mathematics in the 1990 National Assessment of Educational Progress. (ERIC Document Reproduction Service No. ED 378 220)

Kromrey, J. D., & Hines, C. V. (1994). Nonrandomly missing data in multiple regression: An empirical comparison of common missing-data treatments. Educational and Psychology Measurement, 54, 573-593.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. Psychometrika, 2, 151-160.

Laird, N. M. (1988). Missing data in longitudinal studies. Statistics in Medicine, 7, 305-315.

Landerman, L. R., Land, K. C., & Pieper, C. F. (1997). An empirical evaluation of the predictive mean matching method for imputing missing values. Sociological Methods and Research, 26, 3-33.

Little, R. J. A. (1992). Regression with Missing X's: A review. Journal of the American Statistical Association, 87, 1227-1237.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. Journal of the American Statistical Association, 90, 1112-1121.

Little, R. J. A., & Rubin, D. B. (1987). Statistical analysis with missing data. New York: Wiley.

Little, R. J. A., & Rubin, D. B. (1989). The analysis of social science data with missing data. Sociological Methods and Research, 18, 292-326.

Little, R. J. A., & Schenker, N. (1995). Missing data. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), Handbook of statistical modeling for the social and behavioral sciences (pp. 39-75). New York: Plenum.

Longford, N. T. (1994). Models for scoring missing responses to multiple-choice items. (ERIC Document Reproduction Service No. ED 382 650)

Marcoulides, G. A. (1990). An alternative method for estimating variance components in generalizability theory. Psychological Report, 66, 379-386.

Michiels, B., & Molenberghs, G. (1997). Protective estimation of longitudinal categorical data with nonrandom dropout. Communication in Statistics — Theory and Method, 26, 65-94.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. Journal of Educational Statistics, 17, 131-154.

Neal, T., & Nianci, G. (1997). Generating multiple imputations for matrix sampling data analyzed with item response models. Journal of Educational and Behavioral Statistics, 22, 425-445.

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. Journal of Educational Measurement, 31, 200-219.

Peterson, R. A. (1994). A meta-analysis of Cronbach's coefficient alpha. Journal of Consumer Research, 21, 381-391.

Pollard, W. E. (1986). Bayesian statistics for evaluation research: An introduction. Beverly Hills, CA: Sage.

Raaijmakers, Q. A. (1999). Effectiveness of different missing data treatments in surveys with Likert-type data: Introducing the relative mean substitution approach. Educational and Psychological Measurement, 59, 725-728.

Raghunathana, E., & Siscovick, S. (1996). A multiple-imputation analysis of a case-control study of the risk of primary cardiac arrest among pharmacologically treated hypertensives. Applied Statistics, 45, 335-352.

Raymond, M. R. (1987). Missing data in evaluation research. Evaluation and the Health Professions, 9, 395-420.

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. Personnel Psychology, 47, 537-550.

Rubin, D. B. (1976). Inference and missing data. Biometrika, 63, 581-592.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. New York: Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years. Journal of the American Statistical Association, 91, 473-489.

Rubin, D. B., & Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. Journal of the American Statistical Association, 81, 366-374.

Rubin, D. B., & Schenker, N. (1991). Multiple imputation in health-care databases: An overview and some applications. Statistics in Medicine, 10, 585-598.

Schafer, J. L. (1997). Analysis of incomplete multivariate data. New York: Chapman & Hall.

Schafer, J. L. (1999). Multiple imputation: A primer. Statistical Methods in Medical Research, 8, 3-15.

Schafer, J. L., & Olsen, M. K. (1988). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. Multivariate Behavioral Research, 33, 545-571.

Schmidt, F. L., Hunter, J. E., & Urry, V. W. (1976). Statistical power in criterion-related validation studies. Journal of Applied Psychology, 61, 473-485.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). Journal of the American Statistical Association, 82, 528-550.

van Buuren, S., Boshuizen, H. C., & Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. Statistics in Medicine, 18, 681-694.

Wang, C. Y., Anderson, G. L., & Prentice, R. L. (1999). Estimation of the correlation between nutrient intake measures under restricted sampling. Biometrics, 55, 711-717.

Wang, R., Sedransk, J., & Jinn, J. H. (1992). Secondary data analysis when there are missing observations. Journal of the American Statistical Association, 87, 952-961.

Way, W. D., & Reese, C. M. (1991). An investigation of the use of simplified IRT models for scaling and equating the TOEFL test. (ERIC Document Reproduction Service No. ED 395 024)

Xie, F., & Paik, M. C. (1997). Multiple imputation methods for the missing covariates in generalized estimating equation. Biometrics, 53, 1538-1546.

Yamamoto, K. (1995). Estimating the effects of test length and test time on parameter estimation using the HYBRID model. (ERIC Document Reproduction Service No. ED 395 035)

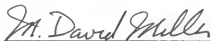
Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. Psychometrika, 52, 275-291.

BIOGRAPHIC SKETCH

Hon Keung Yuen was born in 1961 in Hong Kong. He completed his undergraduate studies at Queensland University, Brisbane, Australia in 1986, where he majored in occupational therapy. In 1988, he received a master of science degree in occupational therapy from Western Michigan University. After five years of occupational therapy practice in the field of traumatic head injury rehabilitation, Mr. Yuen's interest in research grew. Between 1993 and 1996, he taught occupational therapy at Eastern Kentucky University and subsequently in the Hong Kong Polytechnic University. In 1996, he began working on his Ph.D. in the College of Education at the University of Florida, where he majored in research and evaluation methodology.

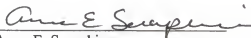
While pursuing his Ph.D., Mr. Yuen also worked full-time on the faculty of the Occupational Therapy Department at the University of Florida. Mr. Yuen has published over twelve articles in the American Journal of Occupational Therapy. Currently, he serves on the editorial board of the American Journal of Occupational Therapy.

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



M. David Miller, Chair
Professor of Educational Psychology

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



Anne E. Seraphine
Assistant Professor of Educational
Psychology

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



Arthur J. Newman
Professor of Educational Leadership, Policy,
and Foundations

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



Kay Walker
Professor of Occupational Therapy

This dissertation was submitted to the Graduate Faculty of the College of Education and the Graduate School and was accepted as partial fulfillment of the requirements for the degree of Doctor of Philosophy.

August, 2000

W. David Miller
Chairman, of Educational Psychology

Ben F. Helms
Dean, College of Education

Dean, Graduate School



.Y942

UNIVERSITY OF FLORIDA



3 1262 08555 1538